The Value of Information in Shortest Path Optimization

by

Michael David Rinehart

S.M., Massachusetts Institute of Technology (2005)
B.S., University of Maryland, College Park (2003)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2010

Author .....
Department of Electrical Engineering and Computer Science
December 2, 2009

Certified by .....
Munther A. Dahleh
Professor
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Theses

The Value of Information in Shortest Path Optimization

by

Michael David Rinehart

Submitted to the Department of Electrical Engineering and Computer Science
on December 2, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Information about a random event (termed the source) is typically treated as a (possibly noisy) function of that event. Information has a destination, an agent, that uses the information to make a decision. In traditional communication systems design, it is usually assumed that the agent uses the information to produce an estimate of the source, and that estimate is in turn used to make the decision. Consequently, the typical objective of communication-systems design is to construct the communication system so that the joint distribution between the source and the information is "optimal" in the sense that it minimizes the average error of the estimate. Due to resource limitations such as cost, power, or time, estimation quality is constrained in the sense that the set of allowable joint distribution is bounded in mutual information.

In the context of an agent using information to make decisions, however, such metrics may not be appropriate. In particular, the true value of information is determined by how it impacts the average payoff of the agent's decisions, not its estimation accuracy. To this end, mutual information may not the most convenient measure of information quantity since its relationship to decision quality may be very complicated, making it difficult to develop algorithms for information optimization.

In this thesis, we study the value of information in an instance of an uncertain decision framework: shortest path optimization on a graph with random edge weights. Specifically, we consider an agent that seeks to traverse the shortest path of a graph subject to some side information it receives about the edge weights in advance of and during its travel. In this setting, decision quality is determined by the average length of the paths the agent chooses, not how often the agent decodes the optimal path.

For this application, we define and quantify a notion of information that is compatible with this problem, bound the performance of the agent subject to a bound on the amount of information available to it, study the impact of spreading information sequentially over partial decisions, and provide algorithms for information optimization. Meaningful, analytic performance bounds and practical algorithms for information optimization are obtained by leveraging a new type of geometric graph reduction for shortest path optimization as well as an abstraction of the geometry of sequential decision making.

Thesis Supervisor: Munther A. Dahleh
Title: Professor

Acknowledgments

First and foremost among those I owe my gratitude is my adviser Prof. Munther Dahleh. I want to first thank him for giving me the opportunity to join his research group and LIDS, the reasons for which of course go without saying. I also wish to thank him for the many lessons I learned from him while a student in his research group, and specifically for what I regard to be the most important of these lessons: how to seek out interesting questions and new landscapes for exploration. While I still have a few more years to go in refining this skill, I am fortunate to start from a such a strong base. Finally, I want to thank him for his excellent teaching and patience, which have contributed significantly to the very positive experience I had while a member of his group.

Within LIDS, I would like to thank some of the current and former students who have made my time here so enjoyable, specifically AmirAli Ahmadi, Aliaa Atwi, Ola Ayaso, Rose Faghih, Sleiman Itani, Sertac Karaman, Yola Katsargyri, Georgios Kotsalis, Patrick Kreidl, Nuno Martins, Sidhant Misra, Paul Njoroge, Mesrob Ohannessian, Mitra Osqui, Mardavij Roozbehani, Keith Santarelli, Sri Sarma, Pari Shah, Noah Stein, Danielle Tarraf, and Ermin Wei. Among these individuals, I want to specifically thank Yola for the many fun conversations, interesting stories, and good times; Paul for the many conversations and good jokes over free CSAIL tea and Muddy Charles beers; and my office mates (Yola, Mitra, and Rose) in part for their willingness to listen to my stories and far-out ideas, but mostly for just for making the office such a fun place to work.

Some friends outside of LIDS that I wish to thank are Al Avestruz, Jose Bohorquez, Wilfried Hofstetter, Deborah Howell, Behnam Jafarpour, Kelly Klima, Bill Nadir, and Crystal Ng. In particular, I would like to thank Al for countless enlightening conversations, good ideas, and just plain good times; Wilfried for many great discussions of geopolitics over many great beers; and Kelly for all the fun times while hanging out at the Thirsty. I also want to thank Lisa Vora and Jason Wampler for the good times and their hospitality whenever I visited Maryland, and I want to thank the many fun and interesting people I met while working with the Thirsty Ear Pub, including the staff (current and former), the executive committee, and the management. My time working with the Thirsty Ear holds a special place in my memories, and everyone involved with the pub is a part of those memories.

Finally, I wish to offer a special set of acknowledgments to my family and my girlfriend.

To my sister Kristin and my brothers Jason, Scott, and Tim, you make going back home more fun than any other destination that I can imagine, and you are collectively funnier than any other crowd I've met. You make traveling back to New Jersey a delight and traveling away a chore.

To my girlfriend Leila, thank you for your love, friendship, personality, sense of humor, support, unending patience, support, and unending patience. I have no doubt that you knew right away that wasn't a typo. There is also no doubt that I would not be here today if it was not for your support of my academic work and aspirations. At the same time, you keep me grounded and well balanced, and you keep me level when stress would otherwise have its own way. You are truly a rock for me as well as my best friend.

And to my parents, I want to thank you for your unending support of my goals at every step of the way. In particular, I want to thank my father for fostering and supporting the development of my interests, skills, and (by example) strong work ethic. I want to thank my mother for her constant belief in me (you were right!) and for her seeming unending wisdom that has had such a profound effect on me. For these reasons and more, I dedicate this dissertation to you both.

Michael Rinehart
November 24, 2009

This work was sponsored by the Army Research Office (W911NF-07-1-0568).

Contents

List of Figures

Chapter 1

Introduction

Information about a random event (termed the *source*) is typically treated as a (possibly noisy) function of that event. Formally, information is as any other random event sharing a joint distribution with the source. Information is considered to have a destination, an *agent*, that uses the information to make a decision. In traditional communication systems design, it is usually assumed that the agent uses the information to produce an estimate of the source, and that estimate is in turn used to make the decision. Consequently, the typical objective of communication systems design is to construct the communication system so that the joint distribution between the source and the information is "optimal" in the sense that it minimizes the average error of the estimate. Due to resource limitations such as cost, power, or time, information optimization is constrained in the sense that the joint distribution is limited to a set of distributions.

A fundamental framework for the study of information communication and estimation error is information theory. In information theory, the total "amount" of information contained in the source is defined as its *entropy*, and the *mutual information* between the source and the information is defined as the amount by which the entropy of the source is reduced on average once we know the information. Bounds on the agent's estimation error can be derived or computed directly from the mutual information. The limitations imposed on optimizing the joint distribution between the source and the information is captured in the abstraction of a *communication channel* over which information is communicated. The communication channel limits the maximum possible mutual information between the source and the agent. This bound is termed the *channel's capacity.*

In the context of an agent using information to make decisions, such traditional metrics may not be appropriate. In particular, the true value of information is determined by how it impacts the average payoff of the agent's decisions, not its estimation quality. To this end, mutual information may not the most convenient measure of information quantity since its relationship to decision quality may be very complicated.

In this thesis, we study the value of information in an uncertain decision framework. We consider an agent who (a) uses information about the source to estimate the quality of its possible decisions (not necessarily the source), (b) can optimize the information it receives

subject to a bound on some useful measure for information quantity, and (c) can consider making partial decisions as information arrives or wait until all of the information arrives before making a complete decision. We specifically explore this framework in the context of shortest path optimization.

In shortest path optimization, an agent uses information about the random edge weights of a graph to determine the shortest-average path in the graph. In this setting, information quality is determined by the average length of the paths the agent chooses, not how often the agent decodes the optimal path. We define and quantify a notion of information that is compatible with this problem, bound the performance of the agent subject to a bound on the amount of information available to it, study the impact of spreading information sequentially over partial decisions, and provide algorithms for optimizing information. Meaningful, analytic performance bounds and practical algorithms for information optimization are obtained by leveraging a new type of geometric graph reduction for shortest path optimization as well as an abstraction partial decision making.

The outline of this thesis is as follows. In Chapter 2, we define the concepts and notations used throughout this thesis, formally describe the framework for shortest path optimization under limited information, and present our objectives. In Chapter 3, we review some research that is, in some cases, qualitatively related to our work and, in other cases, technically related. In Chapter 4, we provide computational and analytic bounds for performance and algorithms for information optimization. In Chapter 5, we study how information spread over partial decisions impacts the agent's performance. In Chapter 6, we study of an alternative but related problem: network flow optimization. And, finally, in Chapter 7, we provide some final conclusions.

Chapter 2

Definitions, Notation, and Objectives

The concepts and tools used in this thesis are based in probability theory, optimization theory, and graph theory, and a working knowledge of these fields is assumed of the reader. In this chapter, we present a set of notation, assumptions, and basic definitions that will be used throughout the remainder of the thesis. A limited set of foundational results for the thesis are also provided.

2.1 Basic Definitions and Notation

2.1.1 Random Variables and Sets

We write random variables (RVs) in capital letters (e.g., X), and denote the event $X \in S$ as $P(X \in S)$ or $P_X(S)$. We write $X \sim p$ if X has p as its probability density function (PDF), and let $N(\mu, \sigma^2)$ be the normal distribution with mean μ and variance σ^2. $\mathrm{E}[X]$ and $\mathrm{VAR}[X]$ are, respectively, the expected value and variance of X, and for a random vector $X = (X_1, \ldots, X_n)$, $\mathrm{VAR}[X] = \sum_{i=1}^{n} \mathrm{VAR}[X_i]$ whereas $\mathrm{COV}[X] = \mathrm{E}[XX^T] - \mathrm{E}[X]\mathrm{E}[X]^T$. For two RVs X, Y, we define $\hat{X}(Y) = \mathrm{E}[X|Y]$ as the estimate of X given Y (which we simplify to $\hat{X}$ if the argument is understood), and we say $X \stackrel{\mathrm{d}}{=} Y$ if both RVs are drawn from the same distribution.

If A is a set, $|A|$ is the number of elements in A. For another set B, $A \backslash B$ is the set of elements in A but not in B. If $A \subset \Re^n$ and $x \in \Re^n$, $A - x = \{a - x \mid a \in A\}$. If A is a subspace, then $A^{\perp}$ is the orthogonal subspace to A, and if A is an affine subspace, then define the subspace $\dot{A} = A - \{a\}$ for any $a \in A$ [1]. We denote the sphere of radius r and center c as $B(r, c) = \{x \mid \|x - c\|^2 \leq r^2\}$.

For a positive (semi)definite matrix M, $N = \sqrt{M}$ is the unique positive (semi)definite matrix satisfying $M = N^2$.

[1] It does not matter which $a \in A$ is selected, $\dot{A}$ will be the same subspace

Finally, for $a \leq b$, let

$$[x]_a^b = \min\Big\{\max\{x, a\}, b\Big\},$$

with $[x]^+ = \max\{x, 0\}$, and $[x]^- = \min\{x, 0\}$. For ease, we simply denote $\mathrm{E}\left[[x]_a^b\right]$ as $\mathrm{E}\,[x]_a^b$.

2.1.2 Graphs and Paths

We define a graph G by a pair (V, E) of vertices V and edges E. Because we allow any two vertices to have multiple edges connect them, we forgo the usual definition $E \subset V \times V$ and instead define a head and tail for each edge $e \in E$ by $\mathrm{hd}\,(e) \in V$ and $\mathrm{tl}\,(e) \in V$ respectively.

Each edge e in the graph is associated with an *edge weight* w_e. The vector of all weights is $w = \left[w_1 \ldots w_{|E|}\right]^T$. Because we consider edge weights to be random, we write the vector as W, and we assume that the probability distribution is known. Finally, we denote the first and second moments of W by

$$\mu = \mathrm{E}\,[W]\,,\ \mu_e = \mathrm{E}\,[W_e]\,,$$
$$\Lambda_W = \mathrm{COV}\,[W]\,,\ \sigma_e^2 = \mathrm{VAR}\,[W_e]\,.$$

We now define the notion of a *path* in the graph.

Definition 1 (Path). *A sequence $p = (e_1, e_2, \ldots, e_n)$ of edges is a path if $\mathrm{tl}\,(e_i) = \mathrm{hd}\,(e_{i+1})$, and we say p goes from $v_1 = \mathrm{hd}\,(e_1)$ to $v_{n+1} = \mathrm{tl}\,(e_n)$.*

Definition 2 (Acyclic Path). *A path $p = (e_i)$ is acyclic if there are no two indices $i < j$ such that $\mathrm{hd}\,(e_i) = \mathrm{tl}\,(e_j)$.*

Assumption 1 (DAG). *G is a directed acyclic graph (DAG) (i.e., all paths the p of G are acyclic).*

We also assume the existence of two vertices $s, t \in V$, respectively termed the *start* and *termination* vertices, that (uniquely) satisfy the following assumption.

Assumption 2. *There is a path from s to each vertex $v \in V \backslash \{s\}$ as well as a path from each vertex $v \in V \backslash \{t\}$ to t.*

Let $P = P(G)$ be the set of all paths from s to t in G. With some abuse of notation, we can write each $p \in P$ as a 0-1 vector in $\Re^{|E|}$ where $p_e = 1$ if $e \in p$ and $p_e = 0$ otherwise. In this case, P is a set of all such vectors in $\Re^{|E|}$. Using our vector notation, the length of a path $p \in P$ is simply $p^T W$.

Let $\mathcal{P} = \text{convex hull}\{P\}$. For an edge weight vector w, the length of the shortest path in the graph can be written as

$$\min_{p \in P}\left\{p^T w\right\} = \min_{p \in \mathcal{P}}\left\{p^T w\right\}.$$

Proposition 1.

$$\mathcal{P} = x \in [0,1]^{|E|} \text{ such that } \sum_{e \in \mathrm{tl}(v)} x_e - \sum_{e \in \mathrm{hd}(v)} x_e = \begin{cases} 1 & v = s \\ -1 & v = t \\ 0 & \text{otherwise} \end{cases}$$

Proof. Let X be the set described in the theorem. It is well known that the shortest path under any set of edge weights w can be expressed as $\min_{p \in X} \{p^T w\}$. Therefore $X = \mathcal{P}$. □

2.2 Partial Information in Stochastic Optimization

To motivate our framework for studying the value of side information in shortest path optimization, we present the basic ideas in a general stochastic optimization context. For the purposes of this section, let W be some RV with a known distribution.

2.2.1 Performance Formulation

Consider the following stochastic optimization:

$$J(W) = \min_{x \in \mathcal{X}} h(x, W).$$

Clearly, since W is a RV, $J(W)$ is also a RV, and so the average performance of the optimization is $\mathrm{E}[J(W)]$.

Consider now the task of finding an "optimal" decision x without having the realization of W. A reasonable objective is to select the x that minimizes the average of the objective:

$$\overline{J} = \min_{x \in \mathcal{X}} \mathrm{E}[h(x, W)].$$

Since $\overline{J}$ is a constant, $\mathrm{E}\left[\overline{J}\right] = \overline{J}$. By Jensen's Inequality, $\mathrm{E}[J(W)] \leq \overline{J}$.

We call the first case (where the realization of W was known) the *full-information* case. We call the latter case the *zero-information* case.

We are interested in formulating an in-between *partial-information* case. To this end, we introduce another RV Y that represents the agent's *side information* about W and write the optimization as a function of our side information:

$$J(Y) = \min_{x \in \mathcal{X}} \mathrm{E}[h(x, W)|Y].$$

Once again, because Y is a RV, $J(Y)$ is also a RV, and so the average performance under Y is simply $\mathrm{E}[J(Y)]$. Clearly, the information Y contains about W is completely determined by their joint-distribution p_{WY}, so we define $J(p_{WY}) = \mathrm{E}[J(Y)]$:

$$J(p_{WY}) = \mathrm{E}\left[\min_{x \in \mathcal{X}} \mathrm{E}[h(x, W)|Y]\right]. \tag{2.1}$$

Remark 1. *Intuitively, we are "averaging-out" the information about $h(x, W)$ that we do not have from Y, much like in the zero-information case. The full- and zero-information cases are easily obtained by substituting $Y = 0$ (a constant) and $Y = W$ to respectively yield $J(Y) = \overline{J}$ and $J(Y) = J(W)$.*

Proposition 2. $\mathrm{E}[J(W)] \leq J(p_{WY}) \leq \overline{J}$

Proof. The proof is a simple application of Jensen's Inequality:

$$\begin{aligned}
\mathrm{E}[J(W)] &= \mathrm{E}\left[\min_{x\in X} h(x, W)\right] = \mathrm{E}\left[\mathrm{E}\left[\min_{x\in X} h(x, W) \,|Y\right]\right] \\
&\leq \mathrm{E}\left[\min_{x\in X} \mathrm{E}[h(x, W)|Y]\right] = J(p_{WY}) \\
&\leq \min_{x\in X} \mathrm{E}[\mathrm{E}[h(x, W)|Y]] = \min_{x\in X} \mathrm{E}[h(x, W)] = \overline{J}.
\end{aligned}$$

□

2.2.2 Quantifying Information and Information Optimization

The agent is also given some flexibility in determining the side information it receives in the form of being able to choose the joint distribution p_{WY}. Without added constraints, though, the agent will choose a distribution that yields $\hat{W}(Y) = W$. Therefore, we define a "bound" Γ to limit the set of allowable distributions. In terms of performance, we seek the solution to

$$J(\Gamma) = \min_{p_{WY}\in\Gamma} J(p_{WY}). \tag{2.2}$$

We call (2.2) the *information optimization.*

To quantify information, we further generalize the concept of an information bound to a family of constraint sets $\{\Gamma(C)\}$ parameterized by a non-negative scalar C called the *capacity.* For ease, we simplify our notation by writing

$$J(C) = J(\Gamma(C))$$

and call $J(C)$ the *optimal performance under capacity C.* Although not critical to the analysis of this thesis, desirable properties of $\Gamma(C)$ include:

- $\Gamma(C_1) \subset \Gamma(C_2)$ if $C_1 \leq C_2$,
- $p_{WY} \in \Gamma(0)$ implies $\mathrm{E}[W|Y] = \mathrm{E}[W]$,
- if $p_{W(Y_1Y_2)} \in \Gamma(C)$, then p_{WY_1} and p_{WY_2} are in $\Gamma(C)$ [2],
- and there exists a C_{max} such that there exist a $p_{WY} \in \Gamma(C_{max})$ satisfying $\hat{W}(Y) = W$.

[2]Note that Y is any side information, and so it can be taken as a tuple of side information as well

2.3 Partial Information in Shortest Path Optimization

We now specialize our framework to shortest path optimization. We begin by defining the information constraint sets $\{\Gamma(C)\}$:

$$\Gamma(C) = \{p_{WY} | \operatorname{VAR}[\operatorname{E}[W|Y]] \leq C\}.$$

Our choice of $\Gamma(C)$ is a practical selection motivated by the analysis that is to follow in this paper. It furthermore obeys our desired properties for $\{\Gamma(C)\}$.

Proposition 3. $0 = \operatorname{VAR}[\operatorname{E}[W]] \leq \operatorname{VAR}[\operatorname{E}[W|Y_1]] \leq \operatorname{VAR}[\operatorname{E}[W|Y_1Y_2]] \leq \operatorname{VAR}[W]$

Proof. The first equality is obvious. The remainder of the inequalities are based on the identity $\operatorname{E}[f(Y)\operatorname{E}[W|Y]] = \operatorname{E}[f(Y)W]$. Without loss of generality, assume $W \in \Re$ and $\operatorname{E}[W] = 0$. For any Y,

$$\begin{aligned}
0 &\leq \operatorname{VAR}[\operatorname{E}[W|Y] - W] \\
&= \operatorname{VAR}[\operatorname{E}[W|Y]] + \operatorname{VAR}[W] - 2\operatorname{E}[\operatorname{E}[W|Y]W] \\
&= \operatorname{VAR}[\operatorname{E}[W|Y]] + \operatorname{VAR}[W] - 2\operatorname{E}\left[\hat{W}(Y)W\right] \\
&= \operatorname{VAR}[\operatorname{E}[W|Y]] + \operatorname{VAR}[W] - 2\operatorname{E}\left[\hat{W}(Y)\operatorname{E}[W|Y]\right] \\
&= \operatorname{VAR}[\operatorname{E}[W|Y]] + \operatorname{VAR}[W] - 2\operatorname{VAR}[\operatorname{E}[W|Y]] \\
&= \operatorname{VAR}[W] - \operatorname{VAR}[\operatorname{E}[W|Y]],
\end{aligned}$$

The remaining inequalities are similarly proved. □

The interpretation of Proposition 3 is that as we add information to our estimate in the form of $Y = (Y_1)$ to $Y = (Y_1, Y_2)$, our measure for information increases. The lower bound represents the case of having zero information, and the upper bound represents the case of having full information.

Now, given a joint distribution p_{WY} between the edge weights W of the graph and the information Y that the agent receives, we can write the agent's average performance as

$$J(p_{WY}) = \operatorname{E}\left[\min_{p \in P}\{p^T \operatorname{E}[W|Y]\}\right], \tag{2.3}$$

which simplifies to

$$J(p_{\hat{W}}) = \operatorname{E}\left[\min_{p \in P}\left\{p^T \hat{W}\right\}\right]. \tag{2.4}$$

Notice that (2.4) only depends on $p_{\hat{W}}$ [3]. We can also equivalently parameterize $\Gamma(C)$ by

$$\Gamma(C) = \left\{p_{\hat{W}} | \operatorname{VAR}\left[\hat{W}\right] \leq C \text{ and there is a } Y \text{ so that } \hat{W} \overset{\mathrm{d}}{=} \operatorname{E}[W|Y]\right\} \tag{2.5}$$

[3] This is true for any linear objective, not just shortest path optimization.

and ignore the joint distribution p_{WY} altogether.

2.3.1 Optional: Shortest Path Optimization under a Mutual Information Bound

Why do we define a family of abstract sets $\{\Gamma(C)\}$ as our information bound and not simply use mutual information? In general, we want to apply bounds $\Gamma(C)$ that yield a nice relationship between C and $J(C)$. For instance, we will see that our variance bounds relate for information relates nicely the shortest path optimization.

We can relate our information bound to mutual information, however. Let

$$\Gamma'(C) = \{p_{WY} | I(W;Y) \leq I_{max}(C)\}$$

where $I_{max}(C) = \max_{p_{WY} \in \Gamma(C)} \{I(W;Y)\}$. Then $\Gamma(C) \subset \Gamma'(C)$ so that $J(\Gamma'(C)) \leq J(C)$.

In general, computing $I_{max}(C)$ is difficult, but so may be computing the performance under mutual information bounds directly. For instance, it is straightforward to see that the sets $\{\Gamma(C)\}$ that we are using for shortest path optimization relate to mutual information via a rate-distortion problem:

$$I_{max}(C) = \min_{\{p_{WY} | \, \mathrm{VAR}[W-\hat{W}] \leq \mathrm{VAR}[W]-C\}} \{I(W;Y)\}.$$

In general, it is not trivial to solve this optimization, especially in the multivariable case.

2.3.2 Specializing to Gaussian Edge Weights

A particular subcase of interest to us is that of Gaussian edge weights. If $Y = W + N$ where W and N are independent, $W \sim N(\mu, \Lambda_W)$ with $\Lambda_W > 0$, and $N \sim N(0, \Lambda_N)$, then

$$\hat{W}(Y) = \Lambda_W \left(\Lambda_W + \Lambda_N\right)^{-1} (Y - \mu) + \mu.$$

Information optimization in this special case is equivalent to designing the distribution of the noise N, or, equivalently, its covariance matrix Λ_N. It is straightforward to show that designing a positive semidefinite Λ_N is equivalent to designing $\Lambda_{\hat{W}}$ (which is equal to $\Lambda_W \left(\Lambda_W + \Lambda_N\right)^{-1} \Lambda_W$) directly subject to several constraints. Denote this new constraint set as $\Gamma_G(C)$:

$$\Gamma_G(C) = \left\{\Lambda_{\hat{W}} \mid 0 \leq \Lambda_{\hat{W}} \leq \Lambda_W \text{ and } \mathrm{Tr}\left(\Lambda_{\hat{W}}\right) \leq C\right\}. \tag{2.6}$$

Note that $\Gamma_G(C)$ is a convex set. It remains convex if we add additional convex constraints such as $\Lambda_{\hat{W}} \sim$ diag.

Finally, for ease, denote $J_G(C) = J(\Gamma_G(C))$.

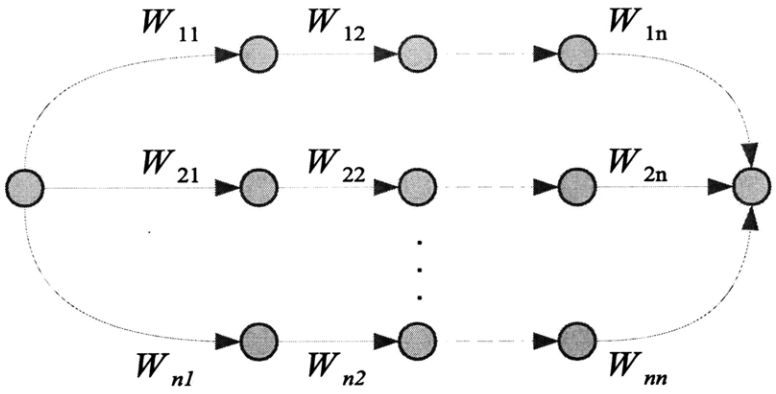


Figure 2-1: Graph for Examples 1 and 2

2.3.3 The Impacts of Graph Topology and Information Selection: Examples

Using our notation from the Gaussian case, we can highlight the impact that even a modest information optimization can have on performance.

Example 1. *Let G be the graph in Figure 2-1 having n disjoint paths $\{p_1, \ldots, p_n\}$ from s to t with each path having n edges each. Let $W_{ij} \sim N(0,1)$ be the (random) weight of edge j on path i and assume the edge weights are independent. Let the "distribution" $\Lambda^s \in \Gamma_G(n)$ satisfy $\Lambda^s \sim$ diagonal, $\Lambda^s_{ee} = 1$ if $e \in p_1$, and $\Lambda^s_{ee} = 0$ otherwise. Essentially, the estimates $\hat{W}$ only contain information about the edges in p_1, meaning that $\hat{W}_e = W_e$ for $e \in p_1$ and $\hat{W}_e = 0$ for $e \notin p_1$.*

For this distribution, the average performance is

$$\begin{aligned} J(\Lambda^s) &= \mathrm{E}\left[\min_i \left\{\sum_j \hat{W}_{ij}\right\}\right] \\ &= \mathrm{E}\left[\min\left\{\sum_j W_{1j}, 0\right\}\right] \\ &= \mathrm{E}\left[\min\left\{\sqrt{n}Z, 0\right\}\right] \\ &= \sqrt{n}\,\mathrm{E}\left[\min\left\{Z, 0\right\}\right] \\ &= -\frac{1}{\sqrt{2\pi}}\sqrt{n} \end{aligned}$$

where $\sum_j W_{1j} \stackrel{\mathrm{d}}{=} \sqrt{n}Z$ with $Z \sim N(0,1)$.

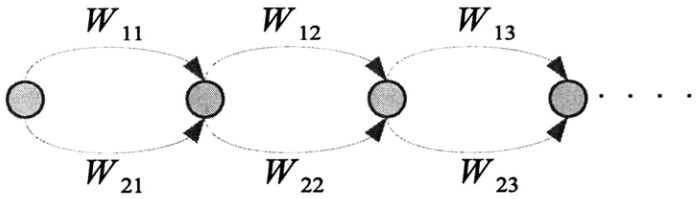


Figure 2-2: Graph for Example 3.

Example 2. *Let G be the graph in Figure 2-1 and take the distribution $\Lambda^p \in \Gamma_G(n)$ where $\Lambda^p \sim$ diagonal, $\Lambda^p_{ee} = 1$ if e is the first edge of any path p_i, and $\Lambda^p_{ee} = 0$ otherwise. Essentially, the estimates $\hat{W}$ only contain information about the first edge in each path, meaning that $\hat{W}_e = W_e$ if e is one of these links and $\hat{W}_e = 0$ otherwise.*

For this distribution, the average performance is

$$\begin{aligned} J(\Lambda^p) &= \mathrm{E}\left[\min_i \left\{\sum_j \hat{W}_{ij}\right\}\right] \\ &= \mathrm{E}\left[\min\{W_{11}+0, W_{21}+0, \ldots, W_{n1}+0\}\right] \\ &= \mathrm{E}\left[\min\{Z_{11}, Z_{21}, \ldots, Z_{n1}\}\right] \\ &\geq -\sqrt{2\ln n} \end{aligned}$$

where the last inequality is obtained by using Lemma 3 in [8].

There is a significant difference between the performances of the two examples. If we increase n, the average performance yielded from applying Λ^s outstrips that obtained using Λ^p quite substantially. This motivate our desire to optimize the information received by the agent.

The next example presents a topology for which the agent's performance significantly improves with capacity.

Example 3. *Consider the graph in Figure 5-5 with $W_{ij} \sim N(0,1)$ and independent and assume $C < \frac{|E|}{2}$. Choose any $\Lambda \in \Gamma_G(C) \bigcap \{\Lambda$ is diagonal$\}$, and denote $\lambda_e = \Lambda_{ee}$. We have*

$$\begin{aligned} J(p_{\hat{W}}) =& \sum_j \mathrm{E}\left[\min\left\{\hat{W}_{1j}, \hat{W}_{2j}\right\}\right] = \sum_j \mathrm{E}\left[\hat{W}_{2j}\right] + \mathrm{E}\left[\min\left\{\hat{W}_{1j} - \hat{W}_{2j}, 0\right\}\right] \\ =& J(0) + \sum_j \min\left\{\sqrt{\lambda_{1j}^2 + \lambda_{2j}^2} Z, 0\right\} = J(0) - \frac{1}{\sqrt{2\pi}} \sum_j \lambda_j, \end{aligned}$$

where $\lambda_j = \sqrt{\lambda_{1j}^2 + \lambda_{2j}^2}$ and $Z \sim N(0,1)$.

The optimal $\{\lambda_j^*\}$ (there are $|E|/2$ of them) are given by the solution to

$$\max_{\lambda} \left\{ \sum_j \lambda_j \right\} \text{ subject to } \sum_j \lambda_j^2 = C$$

yielding $\lambda_j^* = \sqrt{\frac{C}{|E|/2}}$. *Thus,*

$$J_G(C) \leq J(0) - \frac{1}{\sqrt{2\pi}} \frac{|E|}{2} \sqrt{\frac{C}{|E|/2}} = J(0) - \frac{1}{2\sqrt{\pi}} \sqrt{|E|} \sqrt{C}.$$

Example 4. *Finally, we consider a path having a single path from s to t. In this case, we get*

$$J(C) = \mathrm{E}\left[\sum_e \hat{W}_e\right] = \sum_e \mathrm{E}\left[\hat{W}_e\right] = J(0),$$

where $J(0)$ is the performance under no information.

2.4 Sequential Information in Shortest Path Optimization

We now consider the impact of spreading information sequentially over partial decisions, called the *sequential information* framework. This case is quite similar to the previous non-sequential case. We begin by developing the two-stage sequential information framework. The motivation for the formal framework is the following sequence of events:

1. The agent receives some side information Y_1 while it is at vertex s.
2. It estimates the edge weights of the graph using Y_1.
3. It chooses a successor vertex v of s to which to travel.
4. After arriving at v, it receives some side information Y_2.
5. It reestimates the edge weights using Y_1 and Y_2.
6. Finally, it chooses the path starting from v with the smallest expected length.

Remark 2. *A caveat in step 3 is that the agent does not necessarily choose the vertex v that falls along the shortest-path estimated using Y_1. The agent may consider using knowledge of the fact that it will receive future information to influence its choice.*

The side information in the two-stage framework is given by a tuple (Y_1, Y_2) that shares a joint distribution $p_{WY_1Y_2}$ with W. The agent's average performance in this two-stage

setup is

$$J(p_{WY_1Y_2}) = \mathrm{E}\left[\min_{e_1|\,\mathrm{hd}(e_1)=s}\left\{\mathrm{E}\left[W_{e_1}|Y_1\right] + \mathrm{E}\left[\min_{\text{path } p=(e_1,\dots)}\left\{\sum_{e_1\neq e\in p}\mathrm{E}\left[W_e|Y_1Y_2\right]\right\}|Y_1\right]\right\}\right].$$

When making the decision about the first edge e_1 to take, the agent only knows Y_1 as well as its average performance after e_1 subject to receiving more side information. For a given choice e_1, the agent can only take paths p after e_1 (i.e., paths of the form $p = (e_1, \dots)$). It chooses the remainder of p by using its knowledge of both Y_1 and Y_2.

To simplify notation, let $\hat{W}_i = \mathrm{E}\left[W|Y_i\right]$ and $\hat{W}_{12} = \mathrm{E}\left[W|Y_1Y_2\right]$. We can write the two-stage performance more compactly.

Proposition 4.

$$J(p_{WY_1Y_2}) = \mathrm{E}\left[\min_{e|\,\mathrm{hd}(e)=s}\left\{\mathrm{E}\left[\min_{p\in P|p_e=1}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right] \tag{2.7}$$

Proof. First, substitute our $\hat{W}_i$ notation to get

$$J(p_{WY_1Y_2}) = \mathrm{E}\left[\min_{e_1|\,\mathrm{hd}(e_1)=s}\left\{(\hat{W}_1)_{e_1} + \mathrm{E}\left[\min_{\text{path } p=(e_1,\dots)}\left\{\sum_{e_1\neq e\in p}(\hat{W}_{12})_e\right\}|Y_1\right]\right\}\right]$$

Using (a) the relationship $\hat{W}_1 = \mathrm{E}\left[\hat{W}_1|Y_1\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{W}_1|Y_1Y_2\right]|Y_1\right] = \mathrm{E}\left[\hat{W}_{12}|Y_1\right]$, and (b) that $(\hat{W}_1)_{e_1}$ is a constant with respect to the inner minimization, we can pull $(\hat{W}_1)_{e_1} = \mathrm{E}\left[(\hat{W}_{12})_{e_1}|Y_1\right]$ into the inner expectation and minimization to get

$$= \mathrm{E}\left[\min_{e_1|\,\mathrm{hd}(e_1)=s}\left\{\mathrm{E}\left[\min_{\text{path } p=(e_1,\dots)}\left\{\sum_{e\in p}(\hat{W}_{12})_e\right\}|Y_1\right]\right\}\right].$$

Finally, using our vector notation for paths and changing $e_1 \to e$ yields the expression in the claim. □

2.4.1 The n-Stage Framework

We can easily generalize this framework to n stages, but we will need to write the expression recursively to keep it manageable. First, we assume we have a tuple of n RVs $(Y_1, \dots, Y_n)$ that share a known joint-distribution $p_{WY_1\dots Y_n}$ with W. For ease, define

$$Y_{\overline{i}} = (Y_1, \dots, Y_i),$$
$$\hat{W}_{\overline{i}} = \mathrm{E}\left[W|Y_{\overline{i}}\right].$$

Define the iterative sequence of functions $(J_i)_i$ by

$$J_1(Y_1) = \min_{e_i | \, \mathrm{hd}(e_i)=s} \{\mathrm{E}\left[J_2(Y_2, Y_1, e_1)|Y_1\right]\} \tag{2.8}$$

$$J_i(Y_i, Y_{\overline{i-1}}, \{e_1, \ldots, e_{i-1}\}) = \min_{e_i | \, \mathrm{hd}(e_i)=\mathrm{tl}(e_{i-1})} \{\mathrm{E}\left[J_{i+1}(Y_{i+1}, Y_{\overline{i}}, \{e_1, \ldots, e_i\})|Y_{\overline{i}}\right]\} \tag{2.9}$$

$$J_n(Y_n, Y_{\overline{n-1}}, \{e_1, \ldots, e_{n-1}\}) = \min_{p=(e_1,\ldots,e_{n-1},\ldots)} \left\{p^T \hat{W}_{\overline{n}}\right\}. \tag{2.10}$$

The agent's average performance is

$$J(p_{WY_{\overline{n}}}) = \mathrm{E}\left[J_1(Y_1)\right]. \tag{2.11}$$

2.4.2 Sequential Information Framework: Information Constraints

As in the non-sequential information case, we allow the agent to optimize the information it receives. For a tuple of capacities $(C_1, \ldots, C_n)$, we restrict $p_{WY_{\overline{n}}}$ to a set $\Gamma(C_1, \ldots, C_n)$. The agent's optimal performance under $\Gamma(C_1, \ldots, C_n)$ is

$$J(C_1, \ldots, C_n) = \min_{p_{WY_{\overline{n}}} \in \Gamma(C_1,\ldots,C_n)} \{J(p_{WY_{\overline{n}}})\}. \tag{2.12}$$

We will define this set when we explore the sequential information case in more detail in Chapter 5.

2.5 Objectives

The objectives of this thesis are to:

1. provide computational and analytic upper and lower bounds for $J(C)$;
2. develop an algorithm that approximately or suboptimally solves $\min_{p_{WY} \in \Gamma(C)} \{J(p_{WY})\}$;
3. and efficiently compute analytic lower bounds for $J(C_1, \ldots, C_n)$ by contrasting the performance of the sequential-information case to the non-sequential-information case.

We also present a generalization of our results to a multi-agent setting in Chapter 6, providing a limited study of the impact of information on network flow optimization.

2.6 Chapter Summary

We provided the notation, definitions, and basic concepts that will be used throughout this thesis. The specific frameworks for the value of sequential and non-sequential information in shortest path optimization were introduced and motivated by a general framework of partial information in stochastic optimization.

Chapter 3

Related Research

Several areas of study are related to our formulation and objectives, including the fields of order statistics, stochastic optimization, and shortest path optimization on random graphs. In this chapter, we provide a brief review of the existing works in these fields and detail their relationships to our framework.

3.1 Order Statistics and Optimization

3.1.1 Order Statistics

For a list of RVs $X = \{X_1, X_2, \ldots, X_N\}$, we define the *minimum order statistic* as $X_{(1)} = \min_i \{X_i\}$. Under certain conditions, the PDF of $X_{(1)}$ can be computed using the identity presented in the following (well-known) proposition.

Proposition 5. *If $X_i \sim p_i$ and are independent, then $X_{(1)} \sim p$ with*

$$p(x) = \sum_i p_i(x) \prod_{j \neq i} P(X_j \geq x) \tag{3.1}$$

Proof. The event $X_{(1)} > x$ is equivalent to $\{X_i > x \text{ for all i }\}$. The PDF is computed by taking a derivative. □

$\mathrm{E}\left[X_{(1)}\right]$ can be computed directly from this identity when the X_i's are independent. The computation is even simpler in the case of i.i.d. variables. However, our problem of computing $J(p_{\hat{W}})$ is a special case of computing $\mathrm{E}\left[X_{(1)}\right]$ when the X_i's are not generally independent:

$$J(p_{\hat{W}}) = \mathrm{E}\left[X_{(1)}\right] = \mathrm{E}\left[\min_p \{X_p\}\right]$$

where $X_p = \sum_{e \in p} \hat{W}_e$. In the non-i.i.d. case, computing $\mathrm{E}\left[X_{(1)}\right]$ is difficult, and so bounds are typically computed instead. In this section, we review some methods for computing

these bounds and how they apply to computing bounds for $J(p_{\hat{W}})$.

3.1.2 Analytic Bounds based in Order Statistics

Bounds for $\mathrm{E}\left[X_{(1)}\right]$ in the non-i.i.d. case are typically computed using only certain properties of the distribution, such as the means and variances of the individual RVs. One example of such a bound is the one presented in [1] that is computed by treating the min operator as a norm (by converting it to a max operator) and then bounding it by the 2-norm of the vector X. This method uses only first and second moment information about X. However, the bound is not directly amenable to sums of RVs.

An alternative type of bound is presented in [8] using Jensen's Inequality in combination with the moment generating functions for the X_i's, but the method requires that the X_i's be identically distributed, which does not apply to our framework.

A technique for bounding $\mathrm{E}\left[X_{(1)}\right]$ that can be directly generalized to our case is presented in [5]. Though the bound itself is actually a low-complexity optimization, a conservative analytic expression can be derived simply by choosing a convenient suboptimal solution to the optimization.

We now present a bound for $J(C)$ based in a generalization of Theorem 1 in [5].

Corollary 1. $J(C) \geq J(0) - \frac{1}{2}\sqrt{|E|}C$.

Proof. First, we have

$$\begin{aligned}\min_p \left\{p^T\hat{W}\right\} &= \min_p \left\{p^T(\hat{W}-z)+p^Tz\right\} \geq \min_p \left\{\sum_{e\in p}\hat{W}_e - z_e\right\} + \min_p \left\{p^Tz\right\} \\ &\geq \min_p \left\{\sum_{e\in p}[\hat{W}_e - z_e]^-\right\} + \min_p \left\{p^Tz\right\} \geq \sum_{e\in p}[\hat{W}_e - z_e]^- + \min_p \left\{p^Tz\right\}\end{aligned}$$

The bound above is true for all z. Applying Proposition 1 and Theorem 2 in [5] to the above expression, we get

$$J(p_{\hat{W}}) \geq \max_{z_e}\left\{\min_p \sum_{e\in p} z_e - \sum_e \left(\mathrm{E}\left[\hat{W}_e\right] - z_e + \sqrt{(\mathrm{E}\left[\hat{W}_e\right]-z_e)^2 + \mathrm{VAR}\left[\hat{W}_e\right]}\right)\right\}.$$

Setting $z = \mu$ and applying Jensen's Inequality yields the lower bound in the claim. Since the bound is true for each $p_{\hat{W}}$, it is true for the minimizing such distribution and, thus, $J(C)$. Note that $J(0) = \min_p \left\{\sum_{e\in p}\mu_e\right\}$. □

While Corollary 1 provides an analytic lower bound for $J(C)$, the bound is independent of the graph's topology. Moreover, we will find that it can be easily derived from the analytic bounds in Chapter 4, which does include topology information and lends itself to information optimization.

3.1.3 Computational Lower Bounds via the Generalized Chebyshev's Inequality

Accurate bounds for non-i.i.d. order statistics can also be obtained using mathematical programming techniques. The basis for these approaches is a generalization of Chebyshev's Inequality, first presented in [15] and restated below.

Proposition 6 ([15]). *The following optimization over the distribution p_X*

$$\min_{p_X} \mathrm{E}\left[f_0(X)\right] \text{ subject to}$$
$$\mathrm{E}\left[f_i(X)\right] = z_i \text{ for } 1 \leq i \leq m$$

is lower bounded by

$$\max_{y_0,\{y_i\}} \left\{ y_0 + \sum_i y_i z_i \right\} \text{ subject to}$$
$$y_0 + \sum_{i=1}^{m} y_i f_i(x) \leq f_0(x) \text{ for all } x$$

Proof. Take the Lagrangian dual of the first optimization. □

Under certain conditions, there is no gap between the two optimizations in the proposition [15].

We can specialize Proposition 6 to computing a bound for $J(p_{\hat{W}})$ by setting

$$\begin{aligned} f_0(\hat{W}) &= \min_p \left\{ p^T \hat{W} \right\} \\ f_1(\hat{W}) &= \hat{W} \\ f_{e_1e_2}(\hat{W}) &= \left(\hat{W}\hat{W}^T \right)_{e_1e_2} \\ z_1 &= \mu \\ z_{e_1e_2} &= \Lambda_{e_1e_2} \end{aligned}$$

where Λ is the desired covariance matrix of $\hat{W}$. Applying Proposition 6, we get the following result.

Proposition 7.

$$J(p_{\hat{W}}) \geq \max_{y_0,y,Y} \left\{ y_0 + \mu^T y + \mathrm{Tr}\left(Y(\Lambda^2 + \mu\mu^T)\right) \right\} \text{ subject to}$$
$$\begin{bmatrix} Y & \frac{y-p}{2} \\ \frac{y^T-p^T}{2} & y_0 \end{bmatrix} \geq 0 \text{ for all paths } p$$

Proof. Applying Proposition 6 yields a quadratic constraint that can be rewritten as a semi-definite constraint [5]. □

We can compute a bound for $J(C)$ using a similar approach. We begin with the constraints

$$\begin{aligned} f_0(\hat{W}) &= \min_p \left\{ p^T \hat{W} \right\} \\ f_1(\hat{W}) &= \mathrm{E}\left[\hat{W}\right] \\ f_2(\hat{W}) &= \mathrm{Tr}\left(\mathrm{COV}\left[\hat{W}\right]\right) \\ z_1 &= \mu \\ z_2 &= C \end{aligned}$$

as well as an additional constraint $\mathrm{VAR}\left[\hat{W}_e\right] \leq \sigma_e^2$. The inclusion of this inequality requires a slight change the form of the dual, but, for the purposes of this section, we do not expand on this treatment further.

Unfortunately, the dual optimizations for both $J(p_{\hat{W}})$ and $J(C)$ can have an exponential number of constraints (one for each path). A workaround is presented in [4], but to apply this technique, we need to relax our constraints f_i so that we constrain only the marginals distributions $\{p_{\hat{W}_e}\}$. We can apply this technique to the computation of $J(p_{\hat{W}})$ simply by setting

$$\begin{aligned} f_0(\hat{W}) &= \min_p \left\{ p^T \hat{W} \right\} \\ f_1(\hat{W}) &= \mathrm{E}\left[\hat{W}\right] \\ f_e(\hat{W}_e) &= \mathrm{VAR}\left[\hat{W}_e\right] \\ z_1 &= \mu \\ z_e &= (\Lambda)_{ee} \end{aligned}$$

so that only the individual variances of $\hat{W}_e$ are constrained. Of course, this allow the cross-covariances between the $\hat{W}_e$'s to be completely free.

Unfortunately, the capacity constraint $f_2(\hat{W}) = C$ in optimization for $J(C)$ disobeys the requirements on [4], and it cannot be relaxed without the resulting bound losing meaning. Nonetheless, the overall technique in [4] is useful, and we will present an extension of it in Chapter 4 as a comparison against our techniques.

Two drawbacks of the optimization-based approach taken in this section are the lack of analytic expressions for $J(C)$ and potential over-conservativeness. First, it is not immediately clear how one can derive analytic expressions for performance from the dual optimization, and so the impacts of graph topology and capacity on performance remain unclear using these techniques. Second, the expression for $J(C)$ only depends on the first and second marginal moments of the distributions, and the dependencies between the edge weight estimates generated by the optimization may not even be realizable, so it may not be appropriate to use as the basis for an information optimization algorithm.

3.2 Stochastic Optimization

Another related area of research is stochastic optimization, and numerous papers have considered the problem of bounding the improvement one gets from information in stochastic programming:

- [14] computes the value of full information using by subdividing the domain of the uncertainty and applying numeric algorithms over each domain.
- [2] assumes that $h(x, W)$ is concave in W and then leverage the resulting concavity of $J(W)$ to derive a numeric bound, but the bound is worse than that obtained using Jensen's Inequality.
- [10] considers the imapact of partial information where partial information is represented by a signal that offers information about the underlying distribution of the uncertainty, and it is assumed that there are a finite number of such distributions.
- Finally, [3] considers a similar representation of partial information, but it seeks to determine the worst-case performance of the optimization over the unknown distributions.

In all, the related areas of study in stochastic optimization tend not to be specific to any particular application, and thus the bounds are computational, occasionally overly-conservative, and not amenable to an analytic analysis. Because we are considering a specific formulation (shortest path optimization), we seek to provide both computational and analytic bounds that are based in a common framework tied to the underlying structure of the problem.

3.3 Shortest Paths on Random Graphs

A final related area of study is that of shortest path algorithms on random graphs. However, existing bounds in this area tend to be either largely computational or they place heavy assumptions of the problem formulation to simplify analysis.

Two fairly general papers that can be used to generate bounds for $J(C)$ are [19] and [16]. However, the bounds are independent of the graph's topology. In fact, the bound in [16] is equal to a special case of our bound. Thus, neither bound is amenable to information optimization or the goals of this paper.

[18] and [13] compute bounds for performance that depend of topology and the edge weight distribution, but the assumptions in both are fairly heavy. [18] computes the PDF of the shortest path length in a complete graph having integer edge weights, and [13] studies of the average length of the shortest path in a complete graph with uniform edge weights. They do not generalize to our framework.

Finally, [9] computes a bound for arbitrary graphs having arbitrary distributions using dynamic programming that depends of the topology of the graph. The approach (in our case) produces a lower bound for $J(C)$. However, the algorithm is computational, and it is

also not clear how much information about the graph's structure is actually utilized by the algorithm – by bounding performance along subpaths, some information about intersecting paths may be lost.

We will see that methods used in Chapter 4 supply bounds that depend on graph topology, are amenable to information optimization, and produce analytic bounds for performance.

3.4 Chapter Summary

We reviewed several related areas of study that share some overlap with either our problem formulation, our objectives, or the techniques used in this thesis. We found that none of the works mentioned in these areas address our objectives completely or are special cases of our results.

Chapter 4

The Value of Side Information in Shortest Path Optimization

In this chapter, we present new performance bounds for shortest path optimization under partial information that are amenable to both information optimization and analysis. These bounds are based on a new graph reduction that captures limited but significant information about the geometry of the graph's path polytope. The resulting reduction will allow us to optimize information based on the portions of information that are relevant to determining the shortest path in the graph, not its length, and it will also allow us to capture the effects of the graph's topology on performance.

4.1 Properties of the Path Polytope

Performance measures for graph algorithms are typically expressed in terms of limited characteristics of the graph. Such reductions are useful because they may capture a property of a graph that is either intuitive, easy to compute, or amenable for design, and yet have a significant effect on the algorithm's performance. Examples of well-studied graph reductions include graph diameter, minimal cut, number and classifications of cliques, the spectrum of the matrix form of a graph's matrix (adjacency or Laplacian), and so on.

A challenge in information optimization is the evaluation of the objective $J(p_{\hat{W}})$, which is known to be #P-hard [12]. To overcome this difficulty, we will concern our analysis with meaningful upper and lower bounds for $J(p_{\hat{W}})$ that can be efficiently computed. In this section, we present a graph reduction for shortest path optimization that will allow us to derive such bounds, and we will use these bounds to develop information optimization algorithms as well as analytic performance bounds.

4.1.1 Projection Matrix for the Path Polytope

The first property of $\mathcal{P}$ that we explore is critical to the development of our low-complexity information optimization algorithms. First, let $\overline{p}$ be the path of shortest average length,

defined as $\overline{p} = \text{argmin}_{p \in \mathcal{P}} \{p^T \mu\}$. Clearly, $J(0) = \overline{p}^T \mu$.

Theorem 1. *$\mathcal{P} - \overline{p}$ lies in a strict subspace $\mathcal{S}_{\mathcal{P}}$ of $\Re^{|E|}$.*

Proof. Because $\mathcal{P}$ is a polytope, we only need to show that it does not have volume in $\Re^{|E|}$.

Define $E_v = \{e \mid \text{hd}(e) = v\}$ for $v \in V$. First, assume that $|E_s| > 1$. Define $H \in \Re^{|E_s| \times |E|}$ by

$$H_{(i,j)} = \begin{cases} 1, & i = j \in E_s \\ 0, & \text{otherwise.} \end{cases}$$

Effectively, H is a diagonal projection matrix of $\Re^{|E|}$ onto $\Re^{|E_s|}$ in the sense that if we let $\hat{p} = Hp$, then $\hat{p}_e = p_e$ if $e \in E_s$ and is equal to zero otherwise.

By the virtue of G being DAG and s being the unique start vertex, any path p in G must contain exactly one of the edges in E_s. Hence, $HP = \{p \mid p_e = 1$ for exactly one $e \in E_s$ and $p_e = 0$ otherwise$\}$, and therefore $H\mathcal{P}$ is the simplex in $\Re^{|E_s|}$.

The simplex (and, hence, $H\mathcal{P}$) does not have volume in $\Re^{|E_s|}$. Therefore, $\mathcal{P}$ does not contain a hypercube $\mathcal{C}$ of any size in $\Re^{|E|}$ because if it did, $H\mathcal{C}$ would be a set with volume in the simplex. Because $\mathcal{P}$ does not contain any hypercubes, it has no volume.

Now, assume that $|E_s| \leq 1$. If its equals zero, the claim is obviously true. If it equals one, select the first vertex v "after" s with $|E_v| > 1$ and apply the proof above to E_v. If $|E_v| = 1$ for all v (the only remaining case), then one can easily see that there is only one path in G, the vector [1 1 ... 1], which is a single point in $\Re^{|E|}$ and, thus, has no volume. □

Because G is DAG, we can efficiently characterize $\mathcal{S}_{\mathcal{P}}$ by its projection matrix $H_{\mathcal{P}}$, and we can compute $H_{\mathcal{P}}$ is polynomial time. The details of the computation are not critical to the developments of this paper, so we save the details for the appendix. The following theorem formally states our claim.

Theorem 2. *$H_{\mathcal{P}}$ can be computed in polynomial time.*

4.1.2 Outer Spheric Approximation

The remaining portion of our graph reduction concerns the geometry of the boundary for $\mathcal{P}$. We simplify our description of the boundary by relying on low-complexity outer approximations, specifically a sphere. In general, finding the minimal-radius sphere containing a polytope is computationally hard [6], but in the case of the path polytope, it can be computed quite efficiently.

Theorem 3. *The minimal-radius sphere $B(r_o^*, c_o^*)$ containing $\mathcal{P}$ is given by the solution to*

the convex quadratic optimization

$$\begin{aligned} &\min r^2 \ \textit{subject to} \\ &r^2 \geq J(s) + \|c\|^2 \\ &\overline{J}(v) \geq \max_{e|\,\mathrm{hd}(e)=v} \left\{\overline{J}(\mathrm{tl}\,(e)) + (1-2c_e)\right\} \\ &\overline{J}(t) = 0. \end{aligned} \tag{4.1}$$

Proof. Because the extreme points of $\mathcal{P}$ are the paths p of the graph, a necessary and sufficient condition for a sphere with radius r and center c to contain $\mathcal{P}$ is

$$\|p-c\|^2 \leq r^2 \Leftrightarrow p^T p - 2p^T c + c^T c \leq r^2.$$

However, since p is a 0-1 vector, $p^T p = 1^T p$, so we have the equivalent inequality

$$\begin{aligned} & r^2 \geq p^T(1-2c) + \|c\|^2 \text{ for all } p \in P \\ \Leftrightarrow\ & r^2 \geq \max_{p\in P} \left\{p^T(1-2c)\right\} + \|c\|^2 \\ \Leftrightarrow\ & r^2 \geq \max_{p\in \mathcal{P}} \left\{p^T(1-2c)\right\} + \|c\|^2. \end{aligned}$$

The expression $\max_{p\in\mathcal{P}} \left\{p^T(1-2c)\right\}$ equates to computing the longest path in an acyclic graph when the edge weights are given by $(1-2c)$. It is given by $\overline{J}(s)$ in the claim. □

Corollary 2. $\mathcal{P} - c_o^* \subset \mathcal{S}_{\mathcal{P}}$.

Proof. The proof follows from the intuitively obvious (but not proven in this thesis) fact that $c_o^* \in \mathcal{P}$. □

Remark 3. *Analytic bounds for* (4.1) *can be computed by first choosing a (possibly non-optimal) center c, computing the length of the longest path using under the edge weight vector* $(1-2c)$*, and then computing* r*. We will use this strategy in some later examples where we compute analytic performance bounds under a capacity constraint for certain interesting graph topologies.*

4.1.3 Inner Spheric Approximation

Efficient algorithms for generating inner spheric (as well as ellipsoidal) approximations to a polytope are well-known [6], and we do not reproduce these results in this paper. We do note, however, than such algorithms assume that the set to be approximated has volume, which $\mathcal{P}$ does not. A way around this problem is to simply compute the inner approximation strictly within the affine subspace containing $\mathcal{P}$. The details of the computation are not important to the developments of this paper since we will not leverage inner spheric approximations beyond some basic results that are nearly identical to the case of outer spheric approximations.

Similar to the case of an outer spheric approximation, we let r_i^* and c_i^* respectively be the radius and center of the maximal inner sphere contained in $\mathcal{P}$.

4.2 Information Optimization using Graph Reductions

4.2.1 Information Optimization via Upper and Lower Bound Optimization

In this section, we apply the graph reductions of the last section to derive computational bounds for performance as well as information optimization algorithms. Our first bound follows from a direct application of inner and outer spheric approximations.

Lemma 1.

$$(c_o^*)^T\mu - r_o^* \max_{p_{\hat{W}}\in\Gamma(C)} \left\{\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|\right]\right\} \leq J(C) \leq (c_i^*)^T\mu - r_i^* \min_{p_{\hat{W}}\in\Gamma(C)} \left\{\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|\right]\right\}.$$

Proof. We start with the lower bound. Since $c_o^* \in \mathcal{P}$, $H_{\mathcal{P}}\left(\mathcal{P} - c_o^*\right) = \mathcal{P} - c$. Therefore,

$$\begin{aligned}
J(p_{\hat{W}}) \geq & \mathrm{E}\left[\min_{p\in\mathcal{P}-c_o^*} \left\{(p + c_o^*)^T\hat{W}\right\}\right] \\
= & \mathrm{E}\left[\min_{p\in H_{\mathcal{P}}(\mathcal{P}-c_o^*)} \left\{p^T\hat{W}\right\}\right] + (c_o^*)^T\mu \\
= & \mathrm{E}\left[\min_{p\in(\mathcal{P}-c_o^*)} \left\{(H_{\mathcal{P}}p)^T\hat{W}\right\}\right] + (c_o^*)^T\mu \\
= & \mathrm{E}\left[\min_{p\in(\mathcal{P}-c_o^*)} \left\{p^T H_{\mathcal{P}}\hat{W}\right\}\right] + (c_o^*)^T\mu \\
\geq & \mathrm{E}\left[\min_{p\in B(r_o^*,0)} \left\{p^T H_{\mathcal{P}}\hat{W}\right\}\right] + (c_o^*)^T\mu \\
= & -r_o^* \mathrm{E}\left[\|H_{\mathcal{P}}(\hat{W})\|\right] + (c_o^*)^T\mu.
\end{aligned}$$

The upper bound similarly follows. □

Because evaluating $J(p_{\hat{W}})$ is impractical, we can consider optimizing information by optimizing these bounds. Specifically, we seek to maximize $\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|\right]$ over $\Gamma(C)$. The optimization can be performed using stochastic gradient descent.

Remark 4. *An interpretation of Lemma 1 is that we should optimize side information by concentrating the energy of the estimate to the subspace $\mathcal{S}_{\mathcal{P}}$. The component of $\hat{W}$ normal to that subspace (specifically, $(I - H_{\mathcal{P}})\hat{W}$) is lost in the projection. What does this component represent? It is the amount of length in $\hat{W}$ common to all paths. Thus, it only aids in estimating the actual length of the paths. The tangential component $H_{\mathcal{P}}\hat{W}$ contains all of the information for path selection.*

4.2.2 Upper Bound Optimization in the Gaussian Case

In the Gaussian case, upper bound $J(C)$ by a convex optimization. We begin with a useful background result.

Lemma 2. *Let $Z \sim N(0, I)$. The optimization*

$$\max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\}$$

is equivalent to the convex optimization

$$\max_{\Lambda_{\hat{W}} \in \mathcal{L}} \left\{ \mathrm{E}\left[\| \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\}$$

where $\mathcal{L} = \Gamma_G(C) \bigcap \{\Lambda_{\hat{W}} \mid (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0\}$.

Proof. We have

$$\max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\} = \max_{\Lambda_{\hat{W}} \in \Gamma_G(C), \sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}}} \left\{ \mathrm{E}\left[\| \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\}$$

where the equality comes from the fact that $H_{\mathcal{P}} H_{\mathcal{P}} = H_{\mathcal{P}}$.

Since $\mathrm{Range}(\Lambda_{\hat{W}}) = \mathrm{Range}(\sqrt{\Lambda_{\hat{W}}})$, we have the following equivalences for the constraint $\sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}}$:

$$\begin{aligned} &\sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}} \Leftrightarrow \mathrm{Range}(\sqrt{\Lambda_{\hat{W}}}) = \mathcal{S}_{\mathcal{P}} \Leftrightarrow \mathrm{Range}(\Lambda_{\hat{W}}) = \mathcal{S}_{\mathcal{P}} \Leftrightarrow \Lambda_{\hat{W}} = H_{\mathcal{P}} \Lambda_{\hat{W}} \\ &\Leftrightarrow (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0. \end{aligned}$$

Clearly, this constraint is convex over $\Lambda_{\hat{W}}$. Therefore, $\mathcal{L}$ is convex.

Now we prove the objective is concave. The function $f(\Lambda_{\hat{W}}) = z^T \Lambda_{\hat{W}} z = \| \sqrt{\Lambda_{\hat{W}}} z \|^2$ is linear over $\Lambda_{\hat{W}} \geq 0$. Therefore, $\sqrt{f(\Lambda_{\hat{W}})} = \| \sqrt{\Lambda_{\hat{W}}} z \|$ is concave for $\Lambda_{\hat{W}} \geq 0$. By linearity of the expected value operator, the objective is concave. □

We now state a main result of this chapter.

Theorem 4. *If $p^T \mu \leq K$ for some K over all paths p, $J_G(C)$ is upper bounded by the convex optimization*

$$J_G(C) \leq K - r_i^* \max_{\Lambda_{\hat{W}} \in \mathcal{L}} \left\{ \mathrm{E}\left[\| \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\} \tag{4.2}$$

where $Z \sim N(0, I)$, $\mathcal{L} = \Gamma_G(C) \bigcap \{\Lambda_{\hat{W}} \mid (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0\}$.

Proof. If $p^T\mu \leq K$, then

$$\begin{aligned} J_G(p_{\hat{W}}) &= \mathrm{E}\left[\min_p \left\{p^T\hat{W}\right\}\right] \\ &\leq \mathrm{E}\left[\min_p \left\{p^T\hat{W} + (K - p^T\mu)\right\}\right] \\ &= \mathrm{E}\left[\min_p \left\{p^T(\hat{W} - \mu)\right\}\right] + K. \end{aligned}$$

By Corollary 1 and $\hat{W} \stackrel{\mathrm{d}}{=} \left(\sqrt{\Lambda_{\hat{W}}}Z + \mu\right)$, we seek to compute

$$\max_{\Lambda \in \Gamma(C)} \left\{\mathrm{E}\left[\|H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}}Z\|\right]\right\}.$$

The remainder of the proof follows from Lemma 2. □

4.2.3 Lower Bound Optimization in the Gaussian Case

We can also lower bound $J(C)$ by a convex optimization. We begin with some useful background results.

Proposition 8. *For any function f : paths $\rightarrow \Re$ and any set $\overline{\mathcal{P}} \supset \mathcal{P}$,*

$$J(p_{\hat{W}}) \geq \min_{p \in \mathcal{P}} \{f(p)\} + \min_{p \in \overline{\mathcal{P}}} \left\{p^T\hat{W} - f(p)\right\}.$$

Proof. The proof follows immediately from the fact that $\min_x \{a(x) + b(x)\} \geq \min_x \{a(x)\} + \min_x \{b(x)\}$. □

Remark 5. *Proposition 8 provides a generalization of the approach taken in [5] to generate a low-complexity optimization for bounding the mean of the minimum order statistic. Let $\mathcal{X}$ be the simplex in $\Re^n$ and let $W = (W_1, \ldots W_n)$ be a random vector in $\Re^n$. Then the minimum order statistic is given by $\min_i \{W_i\} = \min_{x \in \mathcal{X}} \left\{x^T W\right\}$. We can derive the bound in [5] follows by setting $f(x) = x^T z$ for some vector $z \in \Re^n$, setting $\overline{\mathcal{X}}$ to the unit cube, and then maximizing over z. A similar approach is taken in our previous work [17] for obtaining an analytic lower bound for $J(C)$.*

Corollary 3. $J(C) \geq J(0) - r_o^* \max_{p_{\hat{W}} \in \Gamma(C)} \left\{\mathrm{E}\left[\|H_{\mathcal{P}}(\hat{W} - \mu)\|\right]\right\}$.

Proof. Applying Proposition 8 with $f(p) = p^T\mu$ and $\overline{\mathcal{P}} = \mathcal{P}$ yields

$$\begin{aligned} J(p_{\hat{W}}) &\geq \min_{p \in \mathcal{P}} \left\{p^T\mu\right\} + \mathrm{E}\left[\min_{p \in \mathcal{P}} \left\{p^T(\hat{W} - \mu)\right\}\right] \\ &= J(0) + \mathrm{E}\left[\min_{p \in \mathcal{P}} \left\{p^T(\hat{W} - \mu)\right\}\right]. \end{aligned}$$

We can now apply Lemma 1 using $\hat{W} - \mu$ in place of $\hat{W}$ yielding

$$J(C) \geq J(0) - r_o^* \max_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}}(\hat{W} - \mu) \| \right] \right\}.$$

□

Remark 6. *Note that both Proposition 8 and Corollary 3 are general results and are not specific to the Gaussian case.*

We now state the lower bound result.

Theorem 5. *$J_G(C)$ is lower bounded by the convex optimization*

$$J_G(C) \geq J(0) - r_o^* \max_{\Lambda \in \mathcal{L}} \left\{ \mathrm{E}\left[\| \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\}$$

where $Z \sim N(0, I)$, $\mathcal{L} = \Gamma_G(C) \bigcap \{\Lambda_{\hat{W}} \mid (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0\}$.

Proof. By Corollary 3,

$$J_G(C) \geq J(0) - r_o^* \max_{p_{WY} \in \Gamma(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}}(\hat{W} - \mu) \| \right] \right\}.$$

By Corollary 1 and our parameterization $\hat{W} = \sqrt{\Lambda_{\hat{W}}} Z + \mu$, we seek to compute

$$\max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\}.$$

The remainder of the proof follows from Lemma 2. □

4.2.4 Special Case Analytic Solution for Information Optimization in the Gaussian Case

Theorems 4 and 5 provide convex optimizations for information optimization. Under certain conditions, we can derive analytic expressions for the optimal performance as well as the corresponding optimizing covariance matrix $\Lambda_{\hat{W}}^*$.

Theorem 6. *Let $m = \dim(\mathcal{S}_{\mathcal{P}})$. If C is sufficiently small and the feasible set for $\Lambda_{\hat{W}}$ is $\Gamma_G(C) = \{\Lambda_{\hat{W}} | 0 \leq \Lambda_{\hat{W}} \leq \Lambda_W$ and $\mathrm{Tr}\left(\Lambda_{\hat{W}}\right) \leq C\}$, then*

$$\max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\| H_{\mathcal{P}} \sqrt{\Lambda_{\hat{W}}} Z \| \right] \right\} = \sqrt{\frac{2}{\pi}} \sqrt{C}$$

with the optimizing $\Lambda_{\hat{W}}^* = U \Sigma U^T$ where U is a unitary matrix of the form $U = [x_1 \cdots x_m \; y_{m+1} \cdots y_{|E|}]$ with $\{x_i\}$ forming an orthonormal basis for $\mathcal{S}_{\mathcal{P}}$, and Σ is a diagonal matrix with $\Sigma_{ii} = C$ for some i and $\Sigma_{jj} = 0$ for $j \neq i$.

Proof. First, we can decompose $\Lambda_{\hat{W}} Z = Z^{\|} + Z^{\perp}$ where $H_{\mathcal{P}} Z^{\|} = Z^{\|}$ and $H_{\mathcal{P}} Z^{\perp} = 0$. By orthogonality, $\|\Lambda_{\hat{W}} Z\|^2 = \|Z^{\|}\|^2 + \|Z^{\perp}\|^2$. Therefore,

$$\|H_{\mathcal{P}} \Lambda_{\hat{W}} Z\| = \|Z^{\|}\| = \sqrt{\|\Lambda_{\hat{W}} Z\|^2 - \|Z^{\perp}\|^2} \leq \|\Lambda_{\hat{W}} Z\|.$$

Hence, $\Lambda^*_{\hat{W}}$ must satisfy $H_{\mathcal{P}} \Lambda^*_{\hat{W}} Z = \Lambda^*_{\hat{W}} Z$ in order to maximize the objective. We can parameterize the set of all feasible $\Lambda_{\hat{W}} \in \Gamma_G(C)$ satisfying this constraint as follows. Set $\Lambda_{\hat{W}} = U \Sigma U^T$ where the unitary matrix U has the form

$$U = [x_1 \ \cdots \ x_m \ y_{m+1} \ \cdots \ y_{|E|}]$$

with $\{x_i\}$ forming an orthonormal basis for $\mathcal{S}_{\mathcal{P}}$, and Σ is a diagonal matrix with $\Sigma_{ii} = C_i$ for $i \leq m$, $\Sigma_{ii} = 0$ for $i > m$, and nonnegative C_i satisfying $\sum_i C_i = C$. Optimizing $\Lambda_{\hat{W}}$ is now equivalent to optimizing over the basis $\{x_i\}$ and constants $\{C_i\}$.

Substituting this parameterization for $\Lambda_{\hat{W}}$ yields the optimization

$$\max_{\{x_i\},\{C_i\}} \left\{ \mathrm{E}\left[\| \sum_{i=1}^{m} C_i x_i x_i^T Z \| \right] \right\} = \max_{\{x_i\},\{C_i\}} \left\{ \mathrm{E}\left[\sqrt{\sum_{i=1}^{m} C_i (x_i^T Z)^2} \right] \right\}$$

where equality follows from the fact that the $\{x_i\}$ are an orthonormal basis.

By the concavity of $\sqrt{\cdot}$ and the fact that $\left\{\frac{C_i}{C}\right\}$ is on a simplex, we have

$$= \max_{\{x_i\},\{C_i\}} \left\{ \sqrt{C}\, \mathrm{E}\left[\sqrt{\sum_{i=1}^{m} \left(\frac{C_i}{C}\right) (x_i^T Z)^2} \right] \right\} \leq \max_{\{x_i\},\{C_i\}} \left\{ \sqrt{C} \sum_{i=1}^{m} \left(\frac{C_i}{C}\right) \mathrm{E}\left[\sqrt{(x_i^T Z)^2} \right] \right\}$$

$$= \max_{\{x_i\},\{C_i\}} \left\{ \sqrt{C} \sum_{i=1}^{m} \left(\frac{C_i}{C}\right) K \right\} = \max_{\{x_i\},\{C_i\}} \left\{ \sqrt{C} K \right\} = \sqrt{C} K$$

where $K = \mathrm{E}\left[\sqrt{(n^T Z)^2}\right]$ for any unit vector n (by symmetry, any unit vector n yields the same number, and hence it is the same for each x_i). The upper bound is achieved if we choose any orthonormal basis $\{x_i\}$ and set $C_i = C$ for some i and $C_j = 0$ for all $j \neq i$. Note that the index i needs to be chosen so that $\Lambda_{\hat{W}} \leq \Lambda_W$. We assume C is sufficiently small in the theorem statement so that there exists such a i that satisfies this condition.

Finally, if we choose $n = [1 \ 0 \ \cdots \ 0]^T$, we get $K = \mathrm{E}\,|Z_1|$ where $|Z_1|$ has the folded normal distribution. Hence $K = \sqrt{\frac{2}{\pi}}$. □

Corollary 4. *Under the conditions of Theorem 6, an optimal* $\Lambda^*_{\hat{W}}$ *is* $\frac{C}{\|p_1 - p_2\|^2}(p_1 - p_2)(p_1 - p_2)^T$ *for any two paths* $p_1, p_2 \in P$.

Proof. Using the definitions of Theorem 6, for the basis element x_i associated with the non-zero element of Σ, make $x_i = \frac{p_1 - p_2}{\|p_1 - p_2\|}$ since then $x_i \in \mathcal{S}_{\mathcal{P}}$. □

Remark 7. *Note that if we were to add additional restrictions to $\Gamma_G(C)$ such as $\Lambda_{\hat{W}} \sim$ diagonal, it may not be practical to derive an analytic solution for the optimizing $\Lambda^*_{\hat{W}}$, and so we would be left with executing a convex optimization to solve for it.*

Remark 8. *An interesting property of the optimal $\Lambda^*_{\hat{W}}$ in Theorem 6 is that it concentrates the information to a single singular vector rather than spreading it over many singular vectors. Corollary 4 goes further and tells us that we should simply compare any two paths p_1 and p_2 of the graph over the edges for they do not intersect (the $p_1 - p_2$ expression). This is in qualitative agreement with Examples 1 and 2 where it was shown that concentrating information energy along a single path of a graph is superior to spreading it among many paths.*

4.3 An Analytic Relationship between Capacity and Performance

4.3.1 Performance Lower Bounds

We can manipulate Corollary 3 to provide analytic lower bounds for performance in terms of the capacity C.

Theorem 7. $J(C) \geq J(0) - r_o^* \sqrt{C}$.

Proof. By $\sigma_{\max}(H_{\mathcal{P}}) = 1$ and Jensen's Inequality,

$$\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|\right] \leq \sqrt{\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|^2\right]} \leq \sqrt{\mathrm{E}\left[\|\hat{W}\|^2\right]} \leq \sqrt{C}$$

for all distributions $p_{\hat{W}} \in \Gamma(C)$. Simply apply this inequality to Corollary 3. □

A lesson that can be drawn from Theorem 7 is that if the designer of the graph has some intuition about the relationship between the graph topology and the radius of the minimal sphere, then the designer should seek to maximize the radius. However, if the radius of the minimal sphere is not known and only a rough estimate of the benefit of information is desired, then the following topology-free bound can be applied.

Corollary 5. *A proportionally-tight lower bound for $J(C)$ over all graph topologies is $J(C) \geq J(0) - \frac{1}{2}\sqrt{|E|}\sqrt{C}$.*

Proof. Applying $c = [\frac{1}{2} \ldots \frac{1}{2}]^T$ in the optimization in Theorem 3 sets $r = \frac{1}{2}\sqrt{|E|}$. Tightness is proven in Example 3. □

Remark 9. *The bound in Corollary 5 appears in [17] and [16] using different methods. In [17], it is obtained by bounding $\mathcal{P}$ with the unit cube. In [16], it is obtained using convex majorization of RVs.*

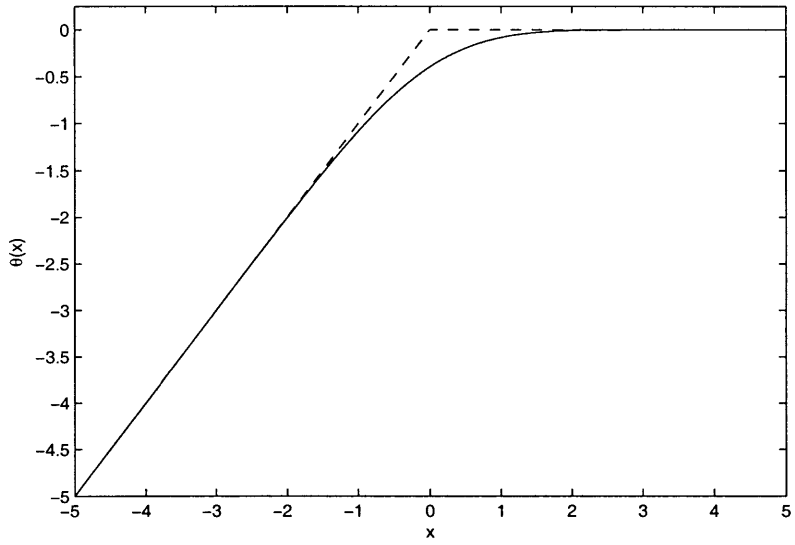


Figure 4-1: $\Theta(x)$ (solid line) compared to $\min\{x, 0\}$ (dashed line).

4.3.2 Performance Upper Bounds

An analytic upper bound for performance is more difficult to provide because the upper bound expressions we have so far derived do not seem to be expressible purely in terms of the capacity C (we cannot apply Jensen's Inequality as we did in the proof of Theorem 7). Moreover, restricted merely to the conditions $\mathrm{E}\left[\hat{W}\right] = \mu$ and $\mathrm{VAR}\left[\hat{W}\right] \leq C$, a tight upper performance bound over distributions for any graph topology is $J(0)$.

Proposition 9. *A tight upper bound for $J(C)$ over all distributions subject to only a mean and variance constraint is $J(0)$ (the zero-information case).*

Proof. Without loss of generality, suppose $\mathrm{E}\left[\hat{W}\right] = 0$ and $\mathrm{VAR}\left[\hat{W}\right] = 1$, and let

$$p_e^n(x) = \frac{1}{2n^2}\delta(x+n) + \frac{1}{2n^2}\delta(x-n) + \frac{2n^2-2}{2n^2}\delta(x).$$

Each distribution p_e^n satisfies the moment constraints, but it is easy to show that as $n \to \infty$, the limiting performance is $J(0)$. □

To this end, we apply a different technique for producing performance upper bounds in terms of C for Gaussian edge weight estimates. The technique is based in reducing $\mathcal{P}$ to a line.

First, let $\Theta_X(c) = \mathrm{E}[\min\{X, c\}]$. Clearly, $\Theta_X(c) \leq c$. Essentially, $\Theta_X(c)$ seeks to determine how much X is less than c on average. When $X \sim N(0,1)$, we will simply write Θ without the subscript. In this case, $\Theta(0) = \frac{-1}{\sqrt{2\pi}}$.

Theorem 8. *Assume independent edge weights and restrict* $\Lambda \in \Gamma_G(C) \bigcap \{\Lambda \text{ is diagonal}\}$. *For any* $\hat{p} \in \mathcal{P}$, *an upper bound for* $J_G(C)$ *is*

$$J_G(C) \leq J(0) + \sqrt{\tilde{C}}\Theta\left(\frac{(\hat{p}-\overline{p})^T\mu}{\sqrt{\tilde{C}}}\right)$$

where $\tilde{C} = \min\left\{C, \text{VAR}\left[(\hat{p}-\overline{p})^T W\right]\right\}$.

Proof. Under independence assumptions, it is easy to verify that

$$\text{VAR}\left[(\hat{p}-\overline{p})^T\hat{W}\right] \leq \min\left\{C, \text{VAR}\left[(\hat{p}-\overline{p})^T W\right]\right\} = \tilde{C}.$$

It is also easy to construct a distribution $p_{\hat{W}}$ that corresponds to a diagonal $\Lambda \in \Gamma_G(C)$ that achieves this bound. Therefore, we have

$$(\hat{p}-\overline{p})^T\hat{W} \stackrel{\text{d}}{=} \sqrt{\tilde{C}}Z + (\hat{p}-\overline{p})^T\mu.$$

where $Z \sim N(0,1)$.

Now,

$$\begin{aligned}
J_G(\Lambda) &= \text{E}\left[\min_{p\in\mathcal{P}}\left\{p^T\hat{W}\right\}\right] \leq \text{E}\left[\min_{p\in\{\overline{p},\hat{p}\}}\left\{p^T\hat{W}\right\}\right] \\
&= \text{E}\left[\min\left\{\overline{p}^T\hat{W}, \hat{p}^T\hat{W}\right\}\right] \\
&= \text{E}\left[\overline{p}^T\hat{W}\right] + \text{E}\left[\min\left\{0, (\hat{p}-\overline{p})^T\hat{W}\right\}\right] \\
&= J(0) + \text{E}\left[\min\left\{(\hat{p}-\overline{p})^T\hat{W}, 0\right\}\right] \\
&= J(0) + \text{E}\left[\min\left\{\sqrt{\tilde{C}}Z + (\hat{p}-\overline{p})^T\mu, 0\right\}\right] \\
&= J(0) + \text{E}\left[\sqrt{\tilde{C}}Z\right] + \text{E}\left[\min\left\{(\hat{p}-\overline{p})^T\mu, -\sqrt{\tilde{C}}Z\right\}\right] \\
&= J(0) + 0 + \sqrt{\tilde{C}}\,\text{E}\left[\min\left\{\frac{(\hat{p}-\overline{p})^T\mu}{\sqrt{\tilde{C}}}, Z\right\}\right].
\end{aligned}$$

□

Corollary 6. *Assume the conditions of Theorem 8 and further assume that* $\hat{p}^T\mu = \overline{p}^T\mu$. *An upper bound for* $J_G(C)$ *is*

$$J_G(C) \leq J(0) - \frac{1}{\sqrt{2\pi}}\sqrt{\tilde{C}}$$

where $\tilde{C} = \min\left\{C, \text{VAR}\left[(\hat{p}-\overline{p})^T W\right]\right\}$.

4.4 Examples

4.4.1 Analytic Examples

We begin with several analytic examples that bound the value of side information under a capacity constraint for certain graph topologies.

Example 5. *In this example, we compute a lower analytic bound for performance for a binary tree with L levels. We first choose a center c for the outer sphere approximation, the process for which corresponds to selecting a center point c_e for each edge e. A binary tree consists of levels of edges, an edge's level being defined as its distance from the root vertex. We select c according to the scheme $c_e = 2^{-l_e}$ where l_e is the level of edge e. Because level l has 2^l edges in it, we have 2^l edges e with $c_e = 2^{-l}$.*

We compute a bound for r_o^* using this center as follows.

$$\begin{aligned}(r_o^*)^2 &\leq \max_p \left\{p^T(1-2c)\right\} + \|c\|^2 \\ &= \left(1 - 2\sum_{l=1}^{L} \frac{1}{2^l}\right) + \left(\sum_l \sum_{e \in layer\ l} 2^l \frac{1}{2^l}\right) \\ &= \frac{1}{2^{L-1}} - 1 + L \\ &\leq L.\end{aligned}$$

Therefore, the average performance under capacity C is lower bounded by $J(C) \geq J(0) - \sqrt{L}\sqrt{C}$.

Example 6. *We now compute a lower analytic bound for performance for a complete graph where the start vertex s and terminating vertex t are chosen randomly. Of course, since a complete graph is undirected, a direct application of (4.1) is not practical since it requires the computation the longest path in an undirected graph with non-negative edge weights, however we can compute a suboptimal analytic lower bound fairly easily.*

For a selection of s and t in the complete graph, we divide the edges of the graph into classes of edges and apply the same center point to each edge within the same class. The classes are

- $\mathcal{A} = \{$*edge connecting s and t*$\}$,
- $\mathcal{B} = \{$*edges connecting s to vertices $v \neq t$*$\}$,
- $\mathcal{C} = \{$*edges connecting $v \neq s$ to t*$\}$,
- *and* $\mathcal{D} = \{$*edges connecting $v \neq s$ to $u \neq t$*$\}$.

We can show that $|\mathcal{A}| = 1$, $|\mathcal{B}| = |\mathcal{C}| = |V| - 2$, and $|\mathcal{D}| = \frac{(|V|-2)(|V|-3)}{2}$.

We assign the center points as follows:

- $e \in \mathcal{A} \Rightarrow c_e = a$,
- $e \in \mathcal{B} \Rightarrow c_e = b$,
- $e \in \mathcal{C} \Rightarrow c_e = c$,
- *and* $e \in \mathcal{D} \Rightarrow c_e = d$.

We can show that as long $0 \leq \{b, c, d\} \leq 1$,

$$r^2 \leq \max\left\{1 - 2a, \max_{2 \leq i \leq |V|-1} \{i - 2(b+c) - (i-2)d\}\right\} + \left(a^2 + (|V|-2)(b^2+c^2) + \frac{(|V|-2)(|V|-3)}{2}d^2\right).$$

Note that $((|V|-1) - 2(b+c) - (|V|-3)d)$ *is the length of the acyclic path having the most edges (that is, the path corresponding to the graph's diameter).*

If we set $a = \frac{1}{2}$ *and* $0 \leq \{b, c, d\} \leq 1$, *we get*

$$r^2 \leq ((|V|-1) - 2(b+c) - (|V|-3)d) + \left(a^2 + (|V|-2)(b^2+c^2) + \frac{(|V|-2)(|V|-3)}{2}d^2\right).$$

Maximizing over b, c, and d, we get $b = c = d = \frac{1}{|V|-2}$, *and*

$$\begin{aligned}(r_o^*)^2 &\leq \frac{4|V|^2 - 13|V| + 4}{4(|V|-2)} \\ &\approx |V| \\ &\approx \sqrt{|E|}\end{aligned}$$

where the final approximate equality is specific to the complete graph. Therefore, the average performance is lower bounded by $J(C) \gtrsim J(0) - \sqrt[4]{|E|}\sqrt{C}$.

Example 7. *Our final analytic example computes a lower bound for average performance improvement on an Erdos-Renyi (E-R) random graph parameterized by* (n, p). *For this example, we assume that the agent is only interested in traveling between connected vertices of the graph, and that it knows the realization of the graph but not the edge weights.*

One can view an E-R graph as a generalization of a complete graph (the complete graph is always achieved if we set $p = 1$*). For some connected component* $\hat{G}$ *of the realization* G, *assign a start vertex* s *and a terminating vertex* t, *use the same classes of edges* $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, *and* $\mathcal{D}$ *as in the complete graph case, and use the same values* a, b, c, *and* d *for those classes applied to the edges in this connected component.*

Suppose there is at least one path from s *to* t *in* $\hat{G}$ *that is not the edge connecting* s *to* t, *and let* $N_{\hat{G}}$ *be the number of edges in the longest such path. Then using a derivation similar*

to that of the complete graph, we get a bound on the radius $r_{\hat{G}}$ for that component:

$$r_{\hat{G}} \leq \sqrt[4]{N_{\hat{G}}}\sqrt{C}.$$

Maximizing this bound (which will serve to lower the performance lower bound) over all components corresponds to maximizing $N_{\hat{G}}$ over $\hat{G}$, the maximum of which is the diameter D_G of the graph.

Therefore, if there is at least one component in G with more than one acyclic path between a pair of two vertices in the component, we have

$$r_o^* \leq \sqrt[4]{D_G}\sqrt{C}.$$

Finally, if $\log n > np \to \infty$, then, as $np \to \infty$, we are guaranteed to have a giant component in the graph, and $D_G \approx \frac{\log n}{\log np}$ [7], yielding

$$\begin{aligned} \mathrm{E}\left[J(C)\right] &\gtrsim \mathrm{E}\left[J(0)\right] - \mathrm{E}\left[\sqrt[4]{D_G}\right]\sqrt{C} \\ &\geq \mathrm{E}\left[J(0)\right] - \sqrt[4]{\mathrm{E}\left[D_G\right]}\sqrt{C} \\ &\approx \mathrm{E}\left[J(0)\right] - \sqrt[4]{\frac{\log n}{\log np}}\sqrt{C}. \end{aligned}$$

4.4.2 Comparative Examples

We now compare our approach for computing performance bounds for $J(C)$ against an optimization designed for the same purpose. The optimization is based in the work of [4] for computing lower bounds for $J(p_{\hat{W}})$. Our extension of [4] will be equivalent to applying an outer approximation $\overline{\Gamma}(C)$ to $\Gamma(C)$ where $\overline{\Gamma}(C)$ is the set of all distributions $p_{\hat{W}}$ satisfying (a) $\mathrm{E}\left[\hat{W}\right] = \mu$, (b) $\mathrm{VAR}\left[\hat{W}_e\right] \leq \sigma_e^2$, and (c) $\mathrm{VAR}\left[\hat{W}\right] \leq C$.

To start, consider the following optimization:

$$\begin{aligned} \underline{J}(C) = \min_{p_{\hat{W}}} &\left\{ \mathrm{E}\left[\min_{p}\left\{p^T\hat{W}\right\}\right]\right\} \text{ subject to} \\ &\mathrm{E}\left[\hat{W}\right] = \mu,\ \mathrm{VAR}\left[\hat{W}\right] \leq C \\ &0 \leq \mathrm{VAR}\left[\hat{W}_e\right] \leq \sigma_e^2. \end{aligned} \tag{4.3}$$

Clearly, $J(C) \geq \underline{J}(C)$. We can solve (4.3) by solving the dual optimization, and there are conditions under which there is no duality gap [15]. However, the dual will have a constraint for each path in the graph, and, hence, may be impractical to solve for even a moderately-sized graph. An efficient approach to tackling optimizations similar to (4.3) is presented in [4], but it requires that the primal only contains constraints on the individual edge weight moments, which our capacity constraint clearly disobeys. Nonetheless, we can

adapt the approach in [4] to handle our capacity constraint quite nicely. We present this extension as a corollary to the main result Corollary 3.2 in [4]. The proof is contained in the appendix.

Corollary 7 (to Corollary 3.2 in [4]).

$$\underline{J}(C) = \min_{\{H_e\},\{\lambda_e\}} \left\{ \sum_e \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \cdot H_e \right\} \text{ subject to}$$

$$H_e \leq 0, \;\; H_e \geq -\begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix}$$

$$\sum_e \left(\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \cdot H_e \right) v_e \in \mathcal{P}$$

$$0 \leq \lambda_e^2 \leq \sigma_e^2, \;\; \sum_e \lambda_e^2 \leq C$$

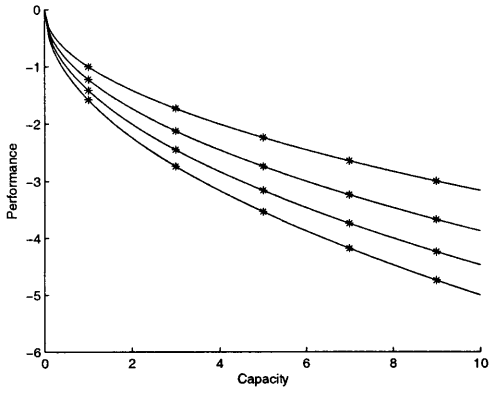
where v_e is the elementary basis vector with the e^{th} component equal to 1.

Remark 10. *While Corollary 7 will yield a tighter performance lower bound than the analytic bound in Theorem 7 (since the Theorem also only uses first and second moment information), it has several drawbacks. First, as we vary C, we need to re-solve the optimization in order to compute the resulting change in $\underline{J}(C)$. Second, it offers no intuition as to how graph topology or capacity impacts performance. Finally, we cannot use it to optimize information in the case of Gaussian edge weights since it only applies limited first and second moment information; it may yield a covariance matrix $\Lambda_{\hat{W}}$ that is not realizable.*

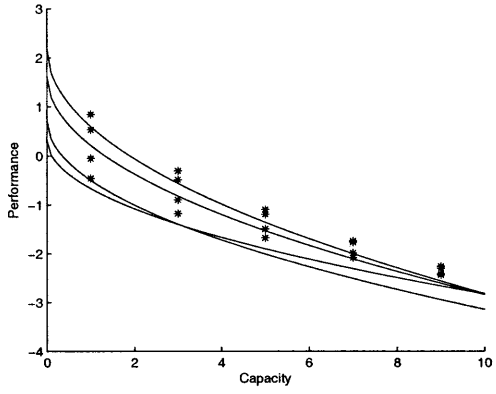
We now present two examples comparing Theorem 7 to Corollary 7. In both, we consider two cases for the mean μ: (a) $\mu = 0$ and (b) a random μ. Because our analytic bounds grow unbounded for increasing C (even for $C > \text{VAR}[W]$), we compare the bounds only over $0 \leq C \leq \min_e \{\text{VAR}[W_e]\}$. This will allow us to examine the impact of increasing capacity without the negative impact of saturating information from any one edge. In both examples, the solution to Corollary 7 is computed using the MATLAB Toolbox CVX [11].

Example 8. *In this example, we consider a graph having two disjoint paths from s to t and consisting of the same number of edges with each edge e satisfying $\text{VAR}[W_e] \geq 10$. We compute performance lower bounds over $0 \leq C \leq 10$. Figures 4-2a and 4-2b show the performances of two approaches in each case. For the case $\mu = 0$, the performances are identical over capacity and graph size. In the case of a random (non-zero) μ, the bound of Corollary 7 has slightly better performance, but it has roughly the same rate of performance improvement.*

Example 9. *In this example, we consider random DAGs consisting of ten vertices and with each edge e satisfying $\text{VAR}[W_e] \geq 10$. We compute performance lower bounds over $0 \leq C \leq 10$. Figures 4-2a and 4-2b show the performances of two approaches in each case.*

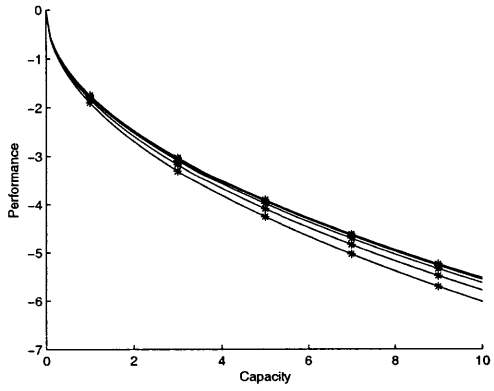


(a) $\mu = 0$ case.

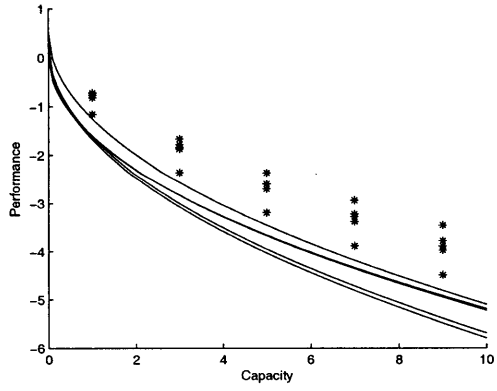


(b) Random μ case

Figure 4-2: The analytic bound of Theorem 7 compared to the optimization-based bound of Corollary 7 for a two-path graph. The solid lines are the analytic bound performances, and the asterisks are the optimization-based bound performances. Each line represents a graph with an increasing number of links per path.



(a) $\mu = 0$ case.



(b) Random μ case

Figure 4-3: The analytic bound of Theorem 7 compared to the optimization-based bound of Corollary 7 for random DAGs. The solid lines are the analytic bound performances, and the asterisks are the optimization-based bound performances. Each line represents a different random graph topology with the number of vertices fixed.

For the case $\mu = 0$, the performances are identical over capacity and graph topology. In the case of a random (non-zero) μ, the bound of Corollary 7 has slightly better performance, but, once again, has roughly the same rate of performance improvement.

There are two interesting observations in the above examples. The first is that the bounds from both Theorem 7 and Corollary 7 seem to be identical in the regime $C \leq \min_e \{\text{VAR}[W_e]\}$ and $\mu = 0$, despite applying different outer-approximations of the constraint set. The second observation is that in the case of non-zero μ, the bounds at least seem to possess the same rate of improvement with increasing capacity.

4.5 Chapter Summary

A new graph reduction for analyzing the value of information for shortest path optimization was presented and used to provide algorithms for information optimization and analytic bounds for performance under a capacity constraint. It was shown that the graph reduction can be efficiently computed, and that information optimization in the case of a Gaussian distribution is bounded by a convex optimization. Examples comparing the analytic bounds of this chapter to an optimization-based bound for the same purpose showed similar (and, in some cases, the same) performance.

Chapter 5

The Value of Sequential Information in Shortest Path Optimization

In this chapter, we consider a generalization of the framework taken in the previous chapter by allowing the agent to receive information as it traverses the graph. We term this setup a *sequential-information* framework, and we consider two special cases of it: *controlled* and *uncontrolled* information. In the controlled information case, the agent can "optimize" the information it receives at each stage based on its location and past information. In the uncontrolled case, the information the agent receives at each stage is independent of the agent's location and past information. The basis for our analysis is an abstraction of sequential decision-making.

5.1 Impact of Applying a Simple Information Constraint Set

Before we can proceed with an analysis of the sequential information cases, we need to define our information constraint set. First, we restate some notation that we first introduced in Chapter 2. For a tuple of capacities $(C_1, \ldots, C_n)$, we restrict $p_{WY_{\overline{n}}}$ to a set $\Gamma(C_1, \ldots, C_n)$. Define

$$\begin{aligned} \hat{W}_i &= \mathrm{E}\left[W|Y_i\right], \\ Y_{\overline{i}} &= (Y_1, \ldots, Y_i), \\ \hat{W}_{\overline{i}} &= \mathrm{E}\left[W|Y_{\overline{i}}\right]. \end{aligned}$$

For the $n = 2$ case, we define $\hat{W}_{12} = \mathrm{E}\left[W|Y_1Y_2\right]$ in order to make $\hat{W}_{\overline{2}} = \hat{W}_{12}$ more discernible $\hat{W}_2$.

The agent's optimal performance under $\Gamma(C_1, \ldots, C_n)$ is

$$J(C_1, \ldots, C_n) = \min_{p_{WY_{\overline{n}}} \in \Gamma(C_1, \ldots, C_n)} \{J(p_{WY_{\overline{n}}})\}. \tag{5.1}$$

Now, one may be tempted to consider the definition $\Gamma(C_1, C_2, \ldots, C_n) = \Gamma(C_1 + C_2 + \cdots + C_n)$, or, equivalently,

$$\left\{ p_{WY_{\overline{n}}} \mid \sum_{i=1}^{n} \mathrm{VAR}\left[\hat{W}_i\right] \leq C \right\}$$

as our information constraint, but this definition has significant drawbacks. In particular, no matter how small (but positive) we make the C_i's, the joint estimate $\hat{W}_{\overline{n}}$ may be arbitrarily accurate in the sense that $\mathrm{VAR}\left[\hat{W}_{\overline{n}}\right] \approx \mathrm{VAR}\left[W\right]$. The following proposition formalizes this claim.

Proposition 10. *For any $\epsilon > 0$,*

$$0 = \inf_{p_{WY_{\overline{n}}}} \mathrm{VAR}\left[W - \hat{W}_{\overline{n}}\right] \quad \textit{subject to}$$
$$\mathrm{VAR}\left[W\right] > 1 - \epsilon, \mathrm{VAR}\left[\hat{W}_i\right] < \epsilon$$

The same bound holds under the restriction $p_{Y_{\overline{n}}} = \prod_i p_{Y_i}$ (independence).

Proof. Without loss of generality, we prove the claim in the case of $n = 2$ and independent Y_1 and Y_2. Define the RV W as a function of the RVs Y_1 and Y_2 by

$$W(Y_1, Y_2) = \begin{cases} h & (Y_1, Y_2) \in \left[0, \frac{1}{h}\right]^2 \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to compute the following quantities:

$$\mathrm{VAR}\left[W\right] = 1 - \frac{1}{h^2}$$
$$\mathrm{VAR}\left[\hat{W}_i\right] = \frac{1}{h} - \frac{1}{h^2} < \frac{1}{h}$$
$$\mathrm{VAR}\left[\hat{W}_{12}\right] = \mathrm{VAR}\left[W\right].$$

Setting h sufficiently large proves the claim. □

An interpretation of Proposition 10 is that regardless of how little the "amount" of side information the agent receives at each time, the total side information can be nearly perfect. However, we are seeking the qualitative relationship "very little side information at each time gives very little total side information." To this end, we will apply assumptions (or limitations) on the structure of the agent's information.

As we already mentioned, we consider two cases of sequential information: controlled and uncontrolled. In the controlled-information case, the agent can "optimize" the information it receives at each stage based on its location and past information. In the uncontrolled case, the information the agent receives at each stage is independent of the agent's location and past information. We will develop different bounds for each of these cases that are partly based in the different definitions for $\Gamma(C_1, \ldots, C_n)$ in each case.

5.2 The Uncontrolled Sequential Information Framework

We begin with an analysis of the uncontrolled information case. We will see that the resulting bound is a generalization of the performance bound in Chapter 4. In particular, if we set $C_1 = C$ and $C_i = 0$ for $i > 1$ (that is, asking for all the side information upfront), we get exactly the analytic lower bound in Chapter 4.

5.2.1 Structuring Information

We begin by defining our information sets $\{\Gamma(C_1, \ldots, C_n)\}$. To motivate the general definition, we start with $n = 2$ and $\mu = 0$:

$$\Gamma(C_1, C_2) = \left\{ p_{WY_1Y_2} \mid \text{VAR}\left[\hat{W}_i\right] \leq C_i \;\&\; \hat{W}_{12} = \hat{W}_1 + \hat{W}_2 \right\}.$$

This definition is reasonable in the sense that, in application, it is desirable to have a low-complexity estimation algorithm. Each time the agent receives information, computing the new estimate is simply a matter of adding the new estimate to the old one. This could correspond to, for example, the Y_i's having information about different edges weights. A consequence of this definition is that $\hat{W}_1$ and $\hat{W}_2$ must be independent.

In the general μ and n case, we want each partial estimation to have a similar structure as in the $n = 2$ case:

$$\Gamma(C_1, \ldots, C_n) = \left\{ p_{WY_{\overline{n}}} \mid \text{VAR}\left[\hat{W}_i\right] \leq C_i \;\&\; \left(\hat{W}_{\overline{k}} - \mu\right) = \sum_{i=1}^{k} \left(\hat{W}_i - \mu\right) \text{ for all } k \leq n \right\}.$$

The proposition below states a consequence of this information structure. We will not use the proposition directly to derive any further results, but it does serve to explicitly address our earlier concern about unbounded VAR $[W_{\overline{n}}]$.

Proposition 11. *For* $p_{WY_{\overline{n}}} \in \Gamma(C_1, \ldots, C_n)$, $\text{VAR}\left[\hat{W}_{\overline{n}}\right] \leq C_1 + \cdots + C_n$.

Proof. Without loss of generality, assume $\mu = 0$. Our constraint set Γ automatically yields the recursive relationship $\hat{W}_{\overline{k}} = \hat{W}_{\overline{k-1}} + \hat{W}_k$. Taking expectations given $Y_{\overline{k-1}}$ yields

$$\text{E}\left[W_{\overline{k}} | Y_{\overline{k-1}}\right] = \hat{W}_{\overline{k-1}} = \hat{W}_{\overline{k-1}} + \text{E}\left[\hat{W}_k | Y_{\overline{k-1}}\right].$$

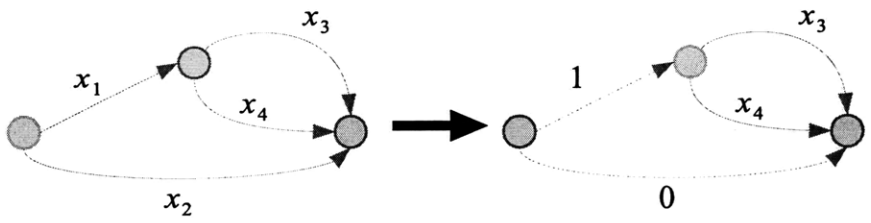


Figure 5-1: Example of a reduction in future decisions based on a past decision.

Hence $\mathrm{E}\left[\hat{W}_k | Y_{\overline{k-1}}\right] = 0$ and so $\hat{W}_k$ is independent of $Y_{\overline{k-1}}$, implying that it is also independent of $\hat{W}_{\overline{k-1}} = \sum_{i=1}^{k-1} \hat{W}_i$. Therefore,

$$\mathrm{VAR}\left[\hat{W}_{\overline{k}}\right] \leq \mathrm{VAR}\left[\hat{W}_{\overline{k-1}}\right] + C_k.$$

Straightforward induction proves that $\mathrm{VAR}\left[\hat{W}_{\overline{k}}\right] \leq \sum_{i=1}^{k} C_i$. □

5.2.2 The Geometry of Partial Decisions in the $n = 2$ Case

A challenge in computing $J(C_1, \ldots, C_n)$ is the evaluation of $J(p_{WY_{\overline{n}}})$, which is known to be #P-hard [12], so, as in Chapter 4, we will instead focus on developing performance bounds for $J(p_{WY_{\overline{n}}})$ and $J(C_1, \ldots, C_n)$. In this section, we develop the geometry underlying partial decisions that will be used to develop such bounds. We motivate our treatment with a simple example.

Example 10. *Consider the left-side graph in Figure 5-1. Any path* $p = [x_1\ x_2\ x_3\ x_4]^T$ *must be one of the vectors*

$$p \in \{[1\ 0\ 1\ 0]^T, [1\ 0\ 0\ 1]^T, [0\ 1\ 0\ 0]^T\} = P.$$

The agent starts at the leftmost vertex and can either choose the top or bottom path. In this example, it chooses the top path. From the next vertex, the set of paths that agent can choose from are

$$p \in \{[1\ 0\ 1\ 0]^T, [1\ 0\ 0\ 1]^T\} = P \bigcap \{[1\ 0\ x_3\ x_4]^T \mid x_i \in \Re\}.$$

The set $\{[1\ 0\ x_3\ x_4]^T \mid x_i \in \Re\}$ *is an affine subspace in* $\Re^4$.

Example 10 highlights the key qualitative concept in the geometry of partial decisions: a partial decision can be represented as an affine subspace S, the intersection of which with the decision space X is the set of remaining possible decisions: $S \bigcap X$.

We now apply this abstraction to our formulation in the two-stage case. We begin with several important definitions. Let

- $E_s = \{e \mid \mathrm{hd}\,(e) = v\}$ be the set of edges connected to s,

- $\hat{E}_e = \{\hat{e} \mid$ there is no path p with $p_e = p_{\hat{e}} = 1\}$ be the set of edges $\hat{e}$ that cannot be reached from e,
- $\overline{E}_e = \{\overline{e} \mid$ there is no path p with $p_e = 1, p_{\overline{e}} = 0\}$ be the set of edges $\overline{e}$ that must be taken if e is taken,
- $S_e = \{x \in \Re^{|E|} \mid x_e = 1, x_{\overline{e}} = 1 \ \forall \overline{e} \in \overline{E}_e, x_{\hat{e}} = 0 \ \forall \hat{e} \in \hat{E}_e\}$ be the affine subspace corresponding to choosing an edge $e \in E_s$,
- and $\mathcal{S} = \{S_e \mid x_e = 1$ for exactly one $e \in E_s\}$ be the set of affine subspaces corresponding to the set of possible decisions at the start vertex s.

We can use our notation to express the two-stage performance in a simpler form.

Lemma 3.

$$J(p_{WY_1Y_2}) = \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right]$$

Proof.

$$\begin{aligned}
&\mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right] \\
=&\mathrm{E}\left[\min_{e\in E_s}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S_e}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right] \\
=&\mathrm{E}\left[\min_{e\in E_s}\left\{\mathrm{E}\left[\min_{p\in P|p_e=1}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right]
\end{aligned}$$

□

5.2.3 Outer Approximating Partial Decisions in the $n = 2$ Case

A key step in our derivation of analytic performance bounds in the uncontrolled case is the use of a relaxation for partial decisions. The relaxation is an outer approximation of the set of affine subspaces corresponding to partial decisions. We begin by motivating the structure of the outer approximation for the two-stage case; that is, outer approximating $\mathcal{S}$.

Our outer approximation consists of larger set of affine subspaces that satisfy a looser set of constraints than $\mathcal{S}$ while maintaining two important properties: (a) the dimension of the maximal-dimensional affine subspace $\max_{S\in\mathcal{S}}\{\dim S\}$, and (b) a particular geometric property of the intersection $S \bigcap B(0,r)$ for $S \in \mathcal{S}$.

We discuss the second property in more detail. A given $S \in \mathcal{S}$ has the form

$$S = \{x \in \Re^{|E|} \mid x_e = 0 \text{ for } e \in E_0, x_e = 1 \text{ for } e \in E_1\}$$

where E_0 and E_1 are appropriately defined (but, of course, do not necessarily satisfy

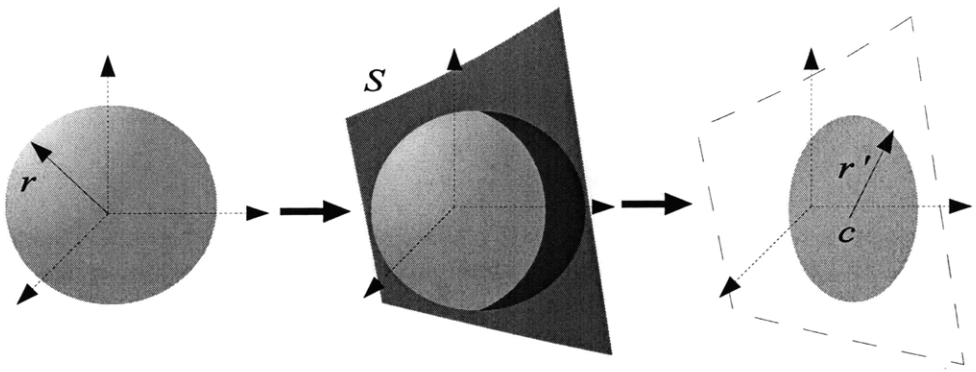


Figure 5-2: Illustration of a 3-dimensional ball with radius r intersecting an affine subspace S of dimension 2. The intersection is a 2-dimensional ball centered at c with radius r' (it appears as an ellipse due to distortion).

$E_0 \bigcup E_1 = E$). The intersection of S with the ball $B(0,r)$ yields a new set:

$$S \bigcap B(r,0) = \left\{ x \in \Re^{|E|} \mid x_e = 0 \text{ for } e \in E_0, x_e = 1 \text{ for } e \in E_1, \sum_{e \notin E_0, E_1} x_e^2 \leq \left(r^2 - \sum_{e \in E_1} 1^2 \right) \right\}.$$

From here, we can show the that $S \bigcap B(0,r) = S \bigcap B(c,r')$ where

$$(c)_e = \begin{cases} 0, & e \notin E_0, E_1 \\ 0, & e \in E_0 \\ 1, & e \in E_1 \end{cases}$$
$$(r')^2 = r^2 - |E_1| = r^2 - \|c\|^2.$$

The key feature to notice is that $r' \leq r$ and $c \in S$. This leads us immediately to our outer approximation for $\mathcal{S}$.

$$\mathcal{S}^m = \{\text{Affine subspaces } S \mid \dim S = m, S \bigcap B(0,r) = S \bigcap B(c,r') \text{ for } r' \leq r, c \in S\}. \tag{5.2}$$

Figure 5-2 illustrates how an affine subspace $S \in \mathcal{S}^2$ operates on a ball in $\Re^3$.

Proposition 12. *If* $\dim S \leq m$ *for every* $S \in \mathcal{S}$, *then*

$$J(p_{WY_1Y_2}) = \mathrm{E}\left[\min_{S \in \mathcal{S}} \left\{ \mathrm{E}\left[\min_{p \in \mathcal{P} \bigcap S} \left\{ p^T \hat{W}_{12} \right\} | Y_1 \right] \right\}\right] \geq \mathrm{E}\left[\min_{S \in \mathcal{S}^m} \left\{ \mathrm{E}\left[\min_{p \in \mathcal{P} \bigcap S} \left\{ p^T \hat{W}_{12} \right\} | Y_1 \right] \right\}\right].$$

Proof. The proof follows from the fact that each $S \in \mathcal{S}$ is contained in some $S' \in \mathcal{S}^m$. $\square$

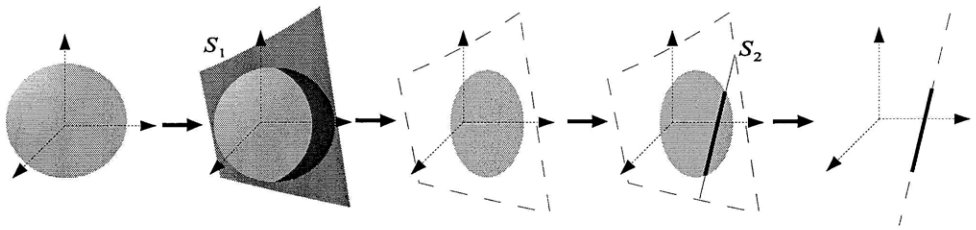


Figure 5-3: Illustration of a 3-dimensional ball intersecting an affine subspace S_1 of dimension 2 and then with another affine subspace $S_2 \subset S_1$ of dimension 1. With each intersection, we get a lower-dimensional ball (a line segment is a ball of dimension 1) with a new radius and center point.

5.2.4 Outer Approximating Partial Decisions in the General n Case

The general n-stage case relaxation follows immediately from the 2-stage case. First, define $\mathcal{S}_1^{m_1} = \mathcal{S}^{m_1}$ to be the set of affine subspaces for the first decision and define the set of affine subspaces for the i^{th} decision as

$$\mathcal{S}_i^{m_i}(S_{i-1}) = \Big\{ \text{Affine subspaces } S_i \mid \\ S_i \subset S_{i-1}, \dim S_i = m_i, S_i \bigcap B(0,r) = S_i \bigcap B(c_{\bar{i}}, r') \text{ for } r' \leq r, c_{\bar{i}} \in S_i \Big\} \tag{5.3}$$

where $S_j \in \mathcal{S}_j^{m_j}(S_{j-1})$ for $j > 1$ and $S_1 \in \mathcal{S}^{m_1}$, and the m_i's are given. Each m_i represents the largest possible dimension of the remaining decision set at the i^{th} step. An important property in this general case is that the i^{th} affine subspace S_i must lie in the affine subspace S_{i-1} of the previous decision. Aside from this modification, the definition of the set is the same as in the two-stage case.

Figure 5-3 illustrates how a sequence of affine subspaces $S_1 \in \mathcal{S}_1^2(\Re^3)$ and $S_2 \in \mathcal{S}_2^1(S_1)$ operate on a ball. Notice that each intersection yields a lower-dimensional ball with a new radius and center. Each successive intersection yields the set of remaining choices available to the agent at that step.

Once again, as an affine subspace, each $S_i \in \mathcal{S}_i$ is completely characterized by (a) a projection matrix H_i (corresponding to its "centered" counterpart $\dot{S}_i$), and (b) an offset vector $c_{\bar{i}} \in S_i$. We select $c_{\bar{i}}$ so that $\|c_{\bar{i}}\| \leq \|x\|$ for all $x \in S_i$. By the Projection Theorem, $c_{\bar{i}}$ is orthogonal to $\dot{S}_i$. It is also intuitively clear that $c_{\bar{i}}$ is the center of the ball resulting from the intersection of S_i with a ball; that is, $S_i \bigcap B(0,r) = S_i \bigcap B(c_{\bar{i}}, r')$. Hence, if we let B_i be the ball resulting from a sequence of intersections with a ball $B(0,r)$ in $\Re^{|E|}$:

$$B_i = B(0,r) \bigcap S_1 \bigcap S_2 \bigcap \cdots \bigcap S_i = B_{i-1} \bigcap S_i$$

with $S_i \in \mathcal{S}(S_{i-1})$, then $c_{\bar{i}}$ is the center of B_i. Of course, $B_0 = \Re^{|E|}$.

Remark 11. *Note that although $c_{\bar{i}}$ is in an affine subspace, the ambient space is still $\Re^{|E|}$, and so we treat it as a vector in $\Re^{|E|}$.*

Of particular interest to us for analysis purposes is the difference $c_{\bar{i}} - c_{\overline{i-1}}$ in the center after each intersection. It will be convenient for us to separate this vector into its length and direction. Thus, we define

$$c_i = \frac{c_{\bar{i}} - c_{\overline{i-1}}}{\|c_{\bar{i}} - c_{\overline{i-1}}\|}$$
$$\gamma_i = \|c_{\bar{i}} - c_{\overline{i-1}}\|.$$

Finally, let

$$\gamma_{\bar{i}} = \|c_{\bar{i}}\|$$

be the distance from the origin to the center of ball B_i.

We now state a series of useful properties regarding these parameters. The first merely states that the center $c_{\bar{i}}$ of ball B_i can be written as the sum of the differences in the centers, which yields a convenient recursion for the centers.

Proposition 13. $c_{\bar{i}} = \sum_{j=1}^{i} \gamma_j c_j = c_{\overline{i-1}} + \gamma_i c_i$.

The second property states that c_i lies in the "centered" subspace $\dot{S}_{i-1}$ but is orthogonal to the new "centered" subspace $\dot{S}_i$.

Proposition 14. $c_i \in \dot{S}_{i-1} \bigcap \dot{S}_i^{\perp}$.

Proof. We prove the claim for $c_i' = \gamma_i c_i = c_{\bar{i}} - c_{\overline{i-1}}$.

Since $S_i \subset S_{i-1}$, $\dot{S}_i \subset \dot{S}_{i-1}$, and, since $c_{\overline{i-1}}$ is orthogonal to $\dot{S}_{i-1}$, it is orthogonal to $\dot{S}_i$ as well. By definition, $c_{\bar{i}}$ is orthogonal to $\dot{S}_i$. Therefore, since c_i' is a linear combination of the two, it is also orthogonal to $\dot{S}_i$.

Choose any $s \in S_{i-1}$. Since $c_{\overline{i-1}} \in S_{i-1}$ and $c_{\bar{i}} \in S_i \subset S_{i-1}$, we have $c_{\overline{i-1}} = x + s$ and $c_{\bar{i}} = y + s$ for some $x, y \in \dot{S}_{i-1}$. Thus, $c_i' = x - y \in \dot{S}_{i-1}$. □

Now we prove an important relationship regarding c_i and the projection matrix H_i. The matrices H_i and $c_i c_i^T$ have orthogonal range spaces (since H_i is the projection matrix for $\dot{S}_i$ and $c_i \perp \dot{S}_i$), but we can capture the range of the sum $H_i + c_i c_i^T$ in the previous subspace $\dot{S}_{i-1}$.

Proposition 15. $H_i + c_i c_i^T \leq H_{i-1}$.

Proof. To prove the claim, we show that $A = \left(H_i + c_i c_i^T\right)$ is a projection matrix for a subspace of $\dot{S}_{i-1}$ since H_{i-1} is a projection matrix for $\dot{S}_{i-1}$.

Since c_i is a unit vector, $c_i c_i^T$ is a projection matrix. Because c_i is orthogonal to $\dot{S}_i$, $c_i^T H_i = [0 \cdots 0]$. It is easy to see that $A^T A = A$ and is thus a projection matrix.

Since H_i is a projection matrix for $\dot{S}_i \subset \dot{S}_{i-1}$, and since $c_i \in \dot{S}_{i-1}$, A is a projection matrix for a subspace of $\dot{S}_{i-1}$. □

Finally, we have an important recursion relating $\gamma_{\bar{i}}$ to $\gamma_{\overline{i-1}}$.

Proposition 16. $\gamma_{\overline{i}} = \sqrt{\gamma_i^2 + \gamma_{\overline{i-1}}^2} = \sqrt{\sum_{j=1}^{i} \gamma_j^2}$.

Proof. By definition,

$$\gamma_i^2 = \gamma_{\overline{i}}^2 + \gamma_{\overline{i-1}}^2 - 2c_{\overline{i}}^T c_{\overline{i-1}}.$$

Therefore, we only need to show that $c_{\overline{i}}^T c_{\overline{i-1}} = \gamma_{\overline{i-1}}^2$. We can write $c_{\overline{i}} = s + c_{\overline{i-1}}$ for some $s \in \dot{S}_{i-1}$. By orthogonality, $s^T c_{\overline{i-1}} = 0$, therefore $c_{\overline{i}}^T c_{\overline{i-1}} = \gamma_{\overline{i-1}}^2$. □

We will heavily leverage the equivalence between S_i and the parameterization $(H_i, c_{\overline{i}})$ in our analysis. For a given sequence of subspaces, we can slightly modify this parameterization to

$$(S_1, S_2, \ldots, S_i) = ((c_i, \gamma_1, H_1), (c_2, \gamma_2, H_2), \ldots, (c_i, \gamma_i, H_i)).$$

5.2.5 Analytic Performance Lower Bound under Uncontrolled Information

We now present the main result for the uncontrolled information case – an analytic performance lower for the uncontrolled information case. Before stating the main result, we state a useful background result that allows us to remove the effect of μ from our bounds and apply our outer approximations.

Lemma 4. $J_1 \geq \underline{J}_1$ *where*

$$\begin{aligned}
\underline{J}_1(Y_1) &= J(0) + \min_{S_1 \in \mathcal{S}_1} \left\{ \mathrm{E}\left[\underline{J}_2(Y_2, Y_1, S_1) | Y_1 \right] \right\} \\
\underline{J}_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) &= \min_{S_i \in \mathcal{S}_i(S_{i-1})} \left\{ \mathrm{E}\left[\underline{J}_{i+1}(Y_{i+1}, Y_{\overline{i}}, S_i) | Y_{\overline{i}} \right] \right\} \\
\underline{J}_n(Y_n, Y_{\overline{n-1}}, S_{n-1}) &= \min_{p \in \overline{\mathcal{P}} \bigcap S_{n-1}} \left\{ p^T \left(\hat{W}_{\overline{n}} - \mu \right) \right\}.
\end{aligned}$$

If $\mu = 0$, *then* $J_i \geq \underline{J}_i$.

Proof. Using the inequality $\min_x \{a(x) + b(x)\} \geq \min_x \{a(x)\} + \min_x \{b(x)\}$, we have

$$\begin{aligned}
J_n((Y_n, Y_{\overline{n-1}}, e_{n-1})) &\geq \min_{p \in P} \left\{ p^T \mu \right\} + \min_{p \in P} \left\{ p^T \hat{W}_{\overline{n}} - p^T \mu \right\} \\
&= J(0) + \min_{p \in P} \left\{ p^T \hat{W}_{\overline{n}} - p^T \mu \right\}.
\end{aligned}$$

It is clear that we can lower bound J_i by applying the outer approximation set $\mathcal{S}_i$. Thus,

we have $J_1 \geq \underline{J}'_1$ where

$$\underline{J}'_1(Y_1) = \min_{S_1 \in \mathcal{S}_1} \left\{ \mathrm{E}\left[\underline{J}'_2(Y_2, Y_1, S_1) | Y_1 \right] \right\}$$
$$\underline{J}'_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) = \min_{S_i \in \mathcal{S}_i(S_{i-1})} \left\{ \mathrm{E}\left[\underline{J}'_{i+1}(Y_{i+1}, Y_{\overline{i}}, S_i) | Y_{\overline{i}} \right] \right\}$$
$$\underline{J}'_n(Y_n, Y_{\overline{n-1}}, S_{n-1}) = J(0) + \min_{p \in \overline{\mathcal{P}} \bigcap S_{n-1}} \left\{ p^T \left(\hat{W}_{\overline{n}} - \mu \right) \right\}.$$

It is straightforward to verify that $\underline{J}_1 = \underline{J}'_1$ and, thus, $J_1 \geq \underline{J}_1$. If $\mu = 0$, it is clear that $J_i \geq \underline{J}_i$. □

The main theorem for the uncontrolled information case is now given.

Theorem 9. *For a given distribution $p_{WY_{\overline{n}}} \in \Gamma(C_1, \ldots, C_n)$ in the uncontrolled-information case, let Δ_i be the sum of the $|E| - m_{i-1}$ smallest singular values of $\mathrm{COV}\left[\hat{W}_i\right]$. For any non-negative capacities $\sum_{i=1}^n C_i = C$,*

$$J(p_{WY_{\overline{n}}}) \geq J(0) - r_o^* \sqrt{C - \sum_{i=1}^n \Delta_i}.$$

Remark 12. *There is a straightforward interpretation of Theorem 9. $|E| - m_{i-1}$ represents the minimum number of edges that either cannot be traversed or must be traversed after the $i^{th} - 1$ stage; namely, those edges for which information at the i^{th} stage is not valuable because either (a) those edges must be taken anyway, or (b) those edges cannot be taken anymore. The effective capacity loss $\sum_{i=1}^n \Delta_i$ indicated in Theorem 9 represents the amount of information energy spread along these edges. If the information is broadcast to the agent or chosen randomly from the graph, an information spread to "useless" edges may be inevitable.*

Proof. The proof is by induction. We will use Proposition 22 in the appendix as the base case. By Lemma 4, we can assume without loss of generality that $\mu = 0$.

First, assume that for any $1 < i < n$,

$$\mathrm{E}\left[J_{i+1}(Y_{i+1}, Y_{\overline{i}}, S_i) \mid Y_{\overline{i}} \right] \geq -\sqrt{(r_o^*)^2 - \gamma_{\overline{i}}^2} \left(\sum_{j=i+1}^n (C_j - \Delta_j) + \hat{W}_{\overline{i}}^T H_i \hat{W}_{\overline{i}} \right)^{\frac{1}{2}} + c_{\overline{i}}^T \hat{W}_{\overline{i}}.$$

We prove that the equivalent bound exists for J_i. To simplify notation, let

$$r^2 = (r_o^*)^2 - \gamma_{\overline{i-1}}^2,$$
$$K_1 = \sum_{j=i+1}^n (C_j - \Delta_j) + \hat{W}_{\overline{i}}^T H_i \hat{W}_{\overline{i}},$$
$$K_2 = c_{\overline{i}}^T \hat{W}_{\overline{i}}.$$

Using the relationship $\gamma_{\bar{i}}^2 = \gamma_i^2 + \gamma_{\overline{i-1}}^2$, we get

$$\mathrm{E}\left[J_{i+1}(Y_{i+1}, Y_{\bar{i}}, S_i) \mid Y_{\bar{i}}\right] \geq -\sqrt{r^2 + \gamma_i^2}\sqrt{K_1} + K_2.$$

Furthermore, since $c_{\bar{i}} = \gamma_i c_i + c_{\overline{i-1}}$, we can expand K_2 as

$$K_2 = \gamma_i c_i^T \hat{W}_{\bar{i}} + c_{\overline{i-1}}^T \hat{W}_{\bar{i}} = \gamma_i c_i^T \hat{W}_{\bar{i}} + K_2',$$

where we select c_i so that $c_i^T \hat{W}_{\bar{i}} \leq 0$ in order to further lower the bound.

By the lower bound in Lemma 4 for J_i, we now have the lower bound

$$\begin{aligned}&\mathrm{E}\left[J_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) \mid Y_{\overline{i-1}}\right] \\ &\quad \geq \mathrm{E}\left[\min_{(\gamma_i, c_i, H_i) \in \mathcal{S}_i(S_{i-1})} \left\{-\sqrt{r^2 + \gamma_i^2}\sqrt{K_1} + \gamma_i c_i^T \hat{W}_{\bar{i}} + K_2'\right\} \mid Y_{\overline{i-1}}\right].\end{aligned}$$

Minimizing over γ_i (actually, over $\gamma_i \geq 0$ if we assume that $c_i^T \hat{W}_{\bar{i}} \leq 0$) yields a new lower bound:

$$\mathrm{E}\left[J_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) \mid Y_{\overline{i-1}}\right] \geq \mathrm{E}\left[\min_{(c_i, H_i) \in \mathcal{S}_i(S_{i-1})} \left\{-\sqrt{r^2}\sqrt{K_1 + (c_i^T \hat{W}_{\bar{i}})^2} + K_2'\right\} \mid Y_{\overline{i-1}}\right].$$

We first examine the term $M = (c_i^T \hat{W}_{\bar{i}})^2$. Note that as M increases, we decrease the lower bound, so we seek an upper bound for M. Using the matrix inequality $c_i c_i^T \leq H_{i-1} - H_i$, we have

$$\begin{aligned}M &= \hat{W}_{\bar{i}} c_i c_i^T \hat{W}_{\bar{i}} \\ &\leq \hat{W}_{\bar{i}} (H_{i-1} - H_i) \hat{W}_{\bar{i}} \\ &= \left(\hat{W}_i H_{i-1} \hat{W}_i + 2\hat{W}_i H_{i-1} \hat{W}_{\overline{i-1}} + \hat{W}_{\overline{i-1}} H_{i-1} \hat{W}_{\overline{i-1}}\right) - \hat{W}_{\bar{i}} H_i \hat{W}_{\bar{i}}\end{aligned}$$

Substituting this upper bound for M yields

$$\begin{aligned}K_1 + M \leq &\left(\sum_{j=i+1}^{n} (C_j - \Delta_j) + \hat{W}_{\bar{i}}^T H_i \hat{W}_{\bar{i}}\right) + \left(\hat{W}_i H_{i-1} \hat{W}_i + 2\hat{W}_i H_{i-1} \hat{W}_{\overline{i-1}} + \hat{W}_{\overline{i-1}} H_{i-1} \hat{W}_{\overline{i-1}}\right) \\ &- \hat{W}_{\bar{i}} H_i \hat{W}_{\bar{i}} \\ = &\sum_{j=i+1}^{n} (C_j - \Delta_j) + \hat{W}_i H_{i-1} \hat{W}_i + 2\hat{W}_i H_{i-1} \hat{W}_{\overline{i-1}} + \hat{W}_{\overline{i-1}} H_{i-1} \hat{W}_{\overline{i-1}} \\ = &K_1'.\end{aligned}$$

Substituting our expressions for K_1' and K_2' yields a new lower bound:

$$\mathrm{E}\left[J_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) \mid Y_{\overline{i-1}}\right] \geq \mathrm{E}\left[\min_{(\gamma_i, c_i, H_i) \in \mathcal{S}_i(S_{i-1})} \left\{-\sqrt{r^2}\sqrt{K_1'}\right\} + K_2' \mid Y_{\overline{i-1}}\right].$$

Because K_1' does not depend on S_i (that is, it is independent of the minimization), we have

$$\begin{aligned}\mathrm{E}\left[J_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) \mid Y_{\overline{i-1}}\right] &\geq \mathrm{E}\left[-\sqrt{r^2}\sqrt{K_1'} + K_2' \mid Y_{\overline{i-1}}\right] \\ &\geq -\sqrt{r^2}\sqrt{\mathrm{E}\left[K_1'|Y_{\overline{i-1}}\right]} + \mathrm{E}\left[K_2'|Y_{\overline{i-1}}\right]\end{aligned}$$

where the inequality comes from Jensen's Inequality.

We now seek to compute the expectations. First,

$$\mathrm{E}\left[K_2'|Y_{\overline{i-1}}\right] = \mathrm{E}\left[c_{\overline{i-1}}^T \hat{W}_{\overline{i}}|Y_{\overline{i-1}}\right] = c_{\overline{i-1}}^T \hat{W}_{\overline{i-1}}.$$

As far as $\mathrm{E}\left[K_1'|Y_{\overline{i-1}}\right]$, we compute an upper bound. Since H_i is a projection matrix for $\dot{S}_i$, $N_i = I - H_i$ is a projection matrix for $\dot{S}_i^{\perp}$. Let $\{n_i^j\}_j$ be an orthonormal basis for $\dot{S}_i^{\perp}$. There are $|E| - m_i$ such vectors, and we can write $N_i = \sum_j n_i^j (n_i^j)^T$. Now, we have

$$\begin{aligned}\mathrm{E}\left[K_1'|Y_{\overline{i-1}}\right] &= \sum_{j=i+1}^{n} (C_j - \Delta_j) + \mathrm{E}\left[\hat{W}_i^T H_{i-1} \hat{W}_i|Y_{\overline{i-1}}\right] + \\ &\quad 2\,\mathrm{E}\left[\hat{W}_i^T|Y_{\overline{i-1}}\right] H_{i-1}\hat{W}_{\overline{i-1}} + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}} \\ &\overset{(a)}{=} \sum_{j=i+1}^{n} (C_j - \Delta_j) + \mathrm{E}\left[\hat{W}_i^T (I - N_{i-1})\hat{W}_i\right] + 0 + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}} \\ &\overset{(b)}{\leq} \sum_{j=i+1}^{n} (C_j - \Delta_j) + C_i - \sum_j (n_j^i)^T \mathrm{E}\left[\hat{W}_i \hat{W}_i^T\right] n_j^i + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}} \\ &\overset{(c)}{\leq} \sum_{j=i+1}^{n} (C_j - \Delta_j) + C_i - \Delta_i + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}} \\ &= \sum_{j=i}^{n} (C_j - \Delta_j) + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}}.\end{aligned}$$

where (a) follows from the independence of $\hat{W}_i$ and $Y_{\overline{i-1}}$ as well as the equality $H = I - N_{i-1}$, (b) follows from $\mathrm{VAR}\left[\hat{W}_i\right] = C_i$ and our expression for N_{i-1} in terms of the vectors n_j^i, and (c) follows from the fact that the n_j^i's are orthonormal vectors.

Therefore,

$$\begin{aligned}&\mathrm{E}\left[J_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) \mid Y_{\overline{i-1}}\right] \\ &\geq -\sqrt{(r_o^*)^2 - \gamma_{\overline{i-1}}^2}\left(\sum_{j=i}^{n} (C_j - \Delta_j) + \hat{W}_{\overline{i-1}}^T H_{i-1}\hat{W}_{\overline{i-1}}\right)^{\frac{1}{2}} + c_{\overline{i-1}}^T \hat{W}_{\overline{i-1}}.\end{aligned}$$

Now we prove the base case. Proposition 22 (see the appendix) provides the lower bound

$$\mathrm{E}\left[J_n(Y_n, Y_{\overline{n-1}}, S_{n-1})|Y_{\overline{n-1}}\right] \geq -\sqrt{(r_o^*)^2 - \gamma_{\overline{n-1}}^2}\left(\mathrm{E}\left[\hat{W}_n^T H_{n-1}\hat{W}_n|Y_{\overline{n-1}}\right]\right.$$
$$\left. + 2\,\mathrm{E}\left[\hat{W}_n^T H_{n-1}\hat{W}_{\overline{n-1}}|Y_{\overline{n-1}}\right] + \hat{W}_{\overline{n-1}}^T H_{n-1}\hat{W}_{\overline{n-1}}\right)^{\frac{1}{2}}$$
$$+ c_{\overline{n-1}}\hat{W}_{\overline{n-1}}.$$

Using an argument similar to that above, we have

$$\mathrm{E}\left[J_n(Y_n, Y_{\overline{n-1}}, S_{n-1})|Y_{\overline{n-1}}\right] \geq -\sqrt{(r_o^*)^2 - \gamma_{\overline{n-1}}^2}\left(C_n - \Delta_n + \hat{W}_{\overline{n-1}}^T H_{n-1}\hat{W}_{\overline{n-1}}\right)^{\frac{1}{2}} + c_{\overline{n-1}}\hat{W}_{\overline{n-1}}.$$

To prove the claim, substitute $S_0 = \Re^{|E|} \Rightarrow H_0 = I \Rightarrow \Delta_1 = 0$, $\hat{W}_0 = \mathrm{E}[W] = 0$, and $\sum_{j=1}^{n} C_i = C$ into the expression for $\mathrm{E}[J_1(Y_1)]$. □

5.3 The Controlled Sequential Information Framework

We now proceed with an analysis of the controlled information case. This case also provides a generalization of the non-sequential case but along different lines. In particular, to make any useful traction in this case, we need to apply a different information structure and performance analysis. For ease, we restrict our discussion in this section to the two-stage case, but it generalizes straightforwardly to the n-stage case. The key development in this section is not the bound itself, but rather practical conditions under which controlled requests may outperform non-sequential requests.

5.3.1 Structuring Information

As in the uncontrolled information case, we seek to structure the information the agent receives so that an accumulation of "little" bit of information at each time does not result in the agent knowing everything about the edge weights. However, we cannot apply the same information restriction used in the uncontrolled case to this case because, in that case, $\hat{W}_1$ and $\hat{W}_2$ were independent. In the controlled case, the agent's location is used to determine future information, but its location is dependent on past information. Thus, there are dependencies among the estimates.

Intuitively, one may expect that the ability for the agent to concentrate future information to the relevant remaining portion of the graph should possibly result in improved performance relative to requesting all information upfront. However, in general, this intuition is not correct. The next proposition shows that if we were to simply constrain information so that $\mathrm{VAR}\left[\hat{W}_{\overline{n}}\right] \leq C_1 + \cdots + C_n$, then it is optimal to request all information upfront.

Proposition 17. *Let $Y = Y_{\overline{n}}$. For any nonnegative capacities C_i, $J(C_1, \ldots, C_n) \geq J(C)$ where $C = \max_{p_{WY} \in \Gamma(C_1,\ldots,C_n)} \left\{\mathrm{VAR}\left[\hat{W}(Y)\right]\right\}$.*

Proof. If $p_{WY_{\overline{n}}} \in \Gamma(C_1, \ldots, C_n)$, then $J(p_{WY_{\overline{n}}}) \geq J(p_{WY})$ by an application of Jensen's Inequality on (2.7). Therefore, since C satisfies $\Gamma(C_1, \ldots, C_n) \subset \Gamma(C)$,

$$\min_{p_{WY_{\overline{n}}} \in \Gamma(C_1,\ldots,C_n)} \{J(p_{WY_{\overline{n}}})\} \geq \min_{p_{WY} \in \Gamma(C_1,\ldots,C_n)} \{J(p_{WY})\} \geq \min_{p_{WY} \in \Gamma(C)} \{J(p_{WY})\}.$$

□

The basis for Proposition 17 is that under such a mild information restriction any joint distribution that can be achieved using controlled requests can also be achieved upfront. This implies that the type of information that the agent can request upfront can be very complicated and can even emulate sequential requests. Hence, we will seek to compare the performance of controlled requests to non-sequential requests in the case where the agent cannot emulate sequential requests upfront. The following example motivates the information structure that we will eventually apply to meet this criterion.

Example 11. *Suppose $W \sim N(0, I)$ and that we restrict the agent to requesting whole edge weights. Therefore, we consider the case $\hat{W}_e \in \{W_e, 0\}$ and integer-valued C_i. We let the agent request C_1 edges weights on the first step, and C_2 different edge weights on the second step.*

Let A_1 be a constant diagonal 0-1 matrix with at most C_1 number of 1's. We have $\hat{W}_1 = A_1 W$. Let A_2 be another constant diagonal 0-1 matrix with at most C_2 number of 1's. The new information the agent gets is $A_2 W$. Because the agent selects different edge weights at the second time step, A_1 and A_2 must have orthogonal range spaces or, equivalently, $A_1^T A_2 = 0$. Finally, because the selection of edge weights at the second step depends on information at the first step, we let A_2 be a function of $\hat{W}_1$ and write $A_2 = A_{2|\hat{W}_1} = A_{2|A_1 W}$. Thus, we can interpret A_2 as a RV.

We cannot write $\hat{W}_2 = A_2 W$ since the very realization of A_2 gives us information about $\hat{W}_1$, and thus $\hat{W}_2$ actually has information outside the range space of A_2. However, we know that our information after the second step is $A_1 W + A_2 W$. Thus, we have $\hat{W}_{12} = A_1 W + A_2 W$. A weaker condition that we satisfy is $\hat{W}_{12} \overset{\mathrm{d}}{=} A_1 W + A_2 W$.

There are two key properties for us in the above example. The first is the orthogonal range space condition $A_1^T A_2 = 0$, and the second is the distribution of the joint estimate $\hat{W}_{12} \overset{\mathrm{d}}{=} A_1 W + A_2 W$. We use these two conditions as the basis for the information structure we apply in the controlled information case.

$$\begin{aligned} \Gamma(C_1, C_2) = \{p_{WY_1Y_2} \mid (\hat{W}_{12} - \mu) &\overset{\mathrm{d}}{=} A_1 Z + A_{2|A_1 Z} Z \\ &\mathrm{VAR}\left[A_1 Z\right] \leq C_1, \mathrm{VAR}\left[A_{2|A_1 Z} Z\right] \leq C_2 \\ &A_1^T A_2 = 0\} \end{aligned}$$

where Z is an RV, A_1 is a constant matrix, and A_2 is a matrix function of $A_1 Z$. From here on, we simplify notation by treating A_2 as a random matrix related to the RV $A_1 Z$.

This information structure limits the types of distributions that agent can use to obtain information, and it guarantees that "little" amounts of information do not accumulate to "too much" information.

Proposition 18. *For* $p_{WY_1Y_2} \in \Gamma(C_1, C_2)$, $\mathrm{VAR}\left[\hat{W}_{12}\right] \leq C_1 + C_2$.

Proof.

$$\begin{aligned}\mathrm{VAR}\left[W_{12}\right] &= \mathrm{VAR}\left[A_1 Z + A_2 Z\right] \\ &= \mathrm{VAR}\left[A_1 Z\right] + \mathrm{VAR}\left[A_2 Z\right] + 2\,\mathrm{E}\left[W^T A_1^T A_2 Z\right] \\ &\leq C_1 + C_2 + 0.\end{aligned}$$

□

5.3.2 Performance Bounds

We now state the main theorem of this section – a comparison of the controlled information case to the non-sequential information case.

Theorem 10. *For* $p_{WY_1Y_2} \in \Gamma(C_1, C_2)$, *a lower bound for the controlled information case is*

$$J(p_{WY_1Y_2}) \geq J(0) - r_o^* \sqrt{C_c},$$

and for $p_{WY} \in \Gamma(C_1 + C_2, 0)$, *a lower bound for the non-sequential information case is*

$$J(p_{WY}) \geq J(0) - r_o^* \sqrt{C_{ns}},$$

where

$$\begin{aligned}C_c &= \max_{A_1} \left\{ \mathrm{E}\left[\max_{A_2} \left\{ \mathrm{VAR}\left[H_{\mathcal{P}}(A_1 Z + A_2 Z) \mid A_1 Z \right] \right\} \right] \right\}, \\ C_{ns} &= \max_{A_1} \left\{ \mathrm{VAR}\left[H_{\mathcal{P}} A_1 Z \right] \right\}.\end{aligned}$$

Furthermore, if A_1^* *is the optimizing matrix for* C_{ns}, *and if there is a projection matrix* H *satisfying* $\mathrm{VAR}\left[H A_1^* Z\right] = C_1$, *then* $C_c \geq C_{ns}$.

Remark 13. *In Chapter 4, it is shown that only the component* $H_{\mathcal{P}}\hat{W}$ *contains useful information about the shortest path in the graph; in other words,* $(I - H_{\mathcal{P}})\hat{W}$ *cannot be used to determine which path is the shortest. Therefore, we should apply all of the information energy into the subspace of the path polytope. One can interpret Theorem 10 as a statement that the agent's performance in the controlled information case can be better than in the non-sequential case because the agent may be better able to concentrate the energy of its estimates to the subspace. This is reflected in the inequality* $C_c \geq C_{ns}$.

Proof. Without loss of generality, assume that $\mathrm{E}[W] = 0$. Write the affine subspace containing $\mathcal{P}$ as $S_{\mathcal{P}} = \dot{S}_{\mathcal{P}} + c_{\mathcal{P}}$. This gives us

$$\begin{aligned}
J(p_{WY_1Y_2}) =& \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{p^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right] \\
\overset{(a)}{=}& \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{(p-c_{\mathcal{P}})^T\hat{W}_{12}\right\}|Y_1\right]\right\}\right] + \mathrm{E}\left[c_{\mathcal{P}}^T W\right] \\
\overset{(b)}{=}& \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{(p-c_{\mathcal{P}})^T H_{\mathcal{P}}\hat{W}_{12}\right\}|Y_1\right]\right\}\right] + 0 \\
\overset{(c)}{=}& \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\mathrm{E}\left[\min_{p\in\mathcal{P}\bigcap S}\left\{p^T H_{\mathcal{P}}\hat{W}_{12}\right\}|Y_1\right]\right\}\right] \\
\overset{(d)}{\geq}& \mathrm{E}\left[\min_{S\in\mathcal{S}}\left\{\min_{p\in\mathcal{P}\bigcap S}\left\{p^T H_{\mathcal{P}}\hat{W}_{12}\right\}\right\}\right] \\
\overset{(e)}{=}& \mathrm{E}\left[\min_{p\in\mathcal{P}}\left\{p^T H_{\mathcal{P}}(A_1Z + A_2Z)\right\}\right]
\end{aligned}$$

where (a) follows from the fact that $c_{\mathcal{P}}$ is a constant independent of p and e_1, (b) follows from the fact that $(p - c_{\mathcal{P}}) \in \dot{\mathcal{S}}_{\mathcal{P}}$ so that $(p - c_{\mathcal{P}}) = H_{\mathcal{P}}(p - c_{\mathcal{P}})$, (c) follows from the fact that $c_{\mathcal{P}}^T H_{\mathcal{P}} = (H_{\mathcal{P}} c_{\mathcal{P}})^T = 0^T$, (d) follows from Jensen's inequality, and (e) follows from substituting the equivalent distribution $\hat{W}_{12} \overset{\mathrm{d}}{=} A_1Z + A_2Z$.

From Theorem 7, we know that $J(C) \geq J(0) - r_o^*\sqrt{\mathrm{VAR}\left[\hat{W}\right]}$. In this case,

$$\hat{W} = H_{\mathcal{P}}(A_1Z + A_2Z).$$

Therefore, for a fixed selection of A_1 and A_2, the agent's performance is

$$J(C_1, C_2) \geq J(0) - r_o^*\sqrt{\mathrm{VAR}\left[H_{\mathcal{P}}(A_1Z + A_2Z)\right]}.$$

Now, in the controlled information case, the agent can optimize A_2 depending on the realization of A_1Z, hence the definition of C_c. In the non-sequential case, the agent can only optimize A_1.

Finally, we prove the final part of the claim. If we set $A_1 = HA_1^*$ and $A_2 = (I - H)A_1^*$, then

$$\begin{aligned}
\mathrm{VAR}\left[A_1Z + A_2Z\right] =& \mathrm{VAR}\left[HA_1^*Z\right] + \mathrm{VAR}\left[(I-H)A_1^*Z\right] + 2\,\mathrm{COV}\left[HA_1^*Z, (I-H)A_1^*Z\right] \\
\leq& C_1 + \left(\mathrm{VAR}\left[A_1^*Z\right] + \mathrm{VAR}\left[HA_1^*Z\right] - 2\,\mathrm{COV}\left[A_1^*Z, HA_1^*Z\right]\right) + 0 \\
\leq& C_1 + (C_1 + C_2) + \mathrm{VAR}\left[HA_1^*Z\right] - 2\,\mathrm{VAR}\left[HA_1^*Z\right] \\
=& C_1 + (C_1 + C_2) - C_1 \\
=& C_1 + C_2,
\end{aligned}$$

and $A_1Z + A_2Z = A_1^*Z$. Therefore, $C_c \geq C_{ns}$. □

Figure 5-4: Graph topology where sequential information is worse than non-sequential information.

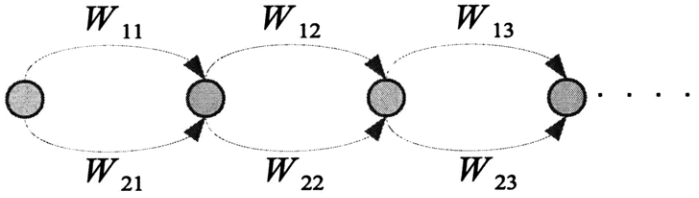


Figure 5-5: Graph topology where (controlled) sequential information can achieve the same performance as the non-sequential case.

5.3.3 Examples

We now present some examples that show how graph topology and our restriction on $\Gamma(C_1, C_2)$ impact whether controlled requests are beneficial relative to non-sequential requests.

Example 12. *Consider the graph in Figure 5-4. The agent can take one of the two edges from the leftmost vertex. After traversing one of those edges, the remainder of its path is determined. Thus, the only useful information about W is contained in the first two edges, and so any information transmitted at the second stage is useless. Regardless of the restrictions on $\Gamma(C_1, C_2)$, $J(C_1, C_2) \geq J(C_1 + C_2)$.*

Example 13. *We reconsider Example 3 from Chapter 2. Consider the graph in Figure 5-5 with $W_{ij} \sim N(0,1)$ and independent. Assume $C < \frac{|E|}{2}$, and suppose we restrict $p_{WY} \in \Gamma(C)$ so that $\hat{W} \sim N(\mu, \Lambda)$ with Λ diagonal (that is, we want our estimates of the edge weights to be independent Gaussians, just like W). In Example 3, it is shown that an optimal such p_{WY} satisfies (a) $\mathrm{VAR}\left[\hat{W}_{1j}\right] = \frac{C}{|E|/2}$ and (b) $\mathrm{VAR}\left[\hat{W}_{2j}\right] = 0$. Of course, by symmetry, one can immediately see that there are at least two such optimal distributions simply by reversing the roles of $\hat{W}_{1j}$ and $\hat{W}_{2j}$.*

It is clear that the agent only requires information about edge weights $\hat{W}_{1j}$ and $\hat{W}_{2j}$ immediately before having to decide which of the two to take (that is, at the j^{th} vertex). For the controlled information case, we take $p_{WY_1Y_2} \in \Gamma(\frac{C}{|E|/2}, C - \frac{C}{|E|/2})$ so that

1. $\mathrm{VAR}\left[\mathrm{E}\left[W_{11}|Y_1\right]\right] = \frac{C}{|E|/2}$ *and* $\mathrm{VAR}\left[\mathrm{E}\left[W_{ij}|Y_1\right]\right] = 0$ *for* $j > 1$,
2. *and* $\mathrm{VAR}\left[\mathrm{E}\left[W_{1j}|Y_1Y_2\right]\right] = \frac{C}{|E|/2}$ *and* $\mathrm{VAR}\left[\mathrm{E}\left[W_{2j}|Y_1Y_2\right]\right] = 0$ *for* $j > 1$.

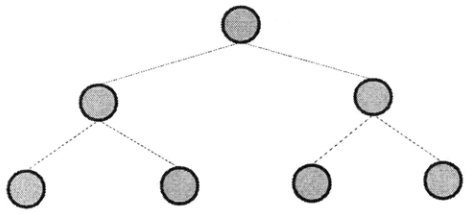


Figure 5-6: Graph topology where controlled sequential information can outperform non-sequential information.

In this case, in is straightforward to see that $J(p_{WY_1Y_2}) = J(C)$. *In fact, we can show that* $J\left(\frac{C}{|E|/2}, C - \frac{C}{|E|/2}\right) = J(C)$, *meaning that* $p_{WY_1Y_2}$ *is optimal.*

To map this setting to our controlled information framework, we set:

- $Z = W$,
- $A_1 \sim$ *diagonal with* $(A_1)_{11} = \sqrt{\frac{C}{|E|/2}}$ *and* $(A_1)_{ii} = 0$ *for* $i > 1$,
- *and* $A_2 \sim$ *diagonal with* $(A_2)_{ii} = \sqrt{\frac{C}{|E|/2}}$ *and* $(A_2)_{11} = 0$.

One can easily check that $\hat{W}_{12} \overset{\mathrm{d}}{=} A_1 W + A_2 W$.

Example 14. *For our final example, we consider the graph in Figure 5-6. In this case, we assume* $W \sim N(0, I)$ *and* $C_1 = C_2 = \frac{1}{2}$. *We restrict* $p_{WY_1Y_2} \in \Gamma(C_1, C_2)$ *so that* $\mathrm{VAR}\left[\mathrm{E}\left[W_e|Y_1\right]\right] \in \left\{0, \frac{1}{2}, 1\right\}$ *and* $\mathrm{VAR}\left[\mathrm{E}\left[W_e|Y_1Y_2\right]\right] \in \left\{0, \frac{1}{2}, 1\right\}$. *In our setup, this is equivalent to setting* $Z = W$ *and restricting* A_1 *and* A_2 *to be diagonal matrices with* $(A)_{ij} \in \left\{0, \sqrt{\frac{1}{2}}, 1\right\}$. *Essentially, we want to restrict the agent to either getting perfect information about one edge or "half" information from each of two edges.*

We now compute the agent's performance under all of the possible request schemes for $p_{WY_1} \in \Gamma(C_1 + C_2, 0) = \Gamma(1, 0)$ *(the non-sequential case):*

Request Scheme	$J(p_{WY_1})$
Perfect information about any one edge	-0.3982
Partial information about two edges coming from the same vertex	-0.3983
Partial information about any two edges in series	-0.4808
Partial information about any two completely disconnected edges	-0.4819

Now, for the controlled information case, we consider only one information scheme. For ease, we simply describe the scheme:

1. *First, request partial information about one of the two first edges coming from the root vertex. Call this edge e.*

2. *If* $\mathrm{E}[W_e|Y_1] \leq 0$, *then traverse* e, *request partial information about one of the two remaining relevant edges, and take the one with the shortest estimated length.*

3. *If* $\mathrm{E}[W_e|Y_1] > 0$, *then traverse the other edge, request partial information about one of the two remaining relevant edges, and take the one with the shortest estimated length.*

Under this information scheme, we get $J(p_{WY_1Y_2}) \approx -0.5658$, *which is better than all of the non-sequential information schemes.*

5.4 Chapter Summary

In this chapter, we compared two types of sequential information schemes, controlled and uncontrolled, to the non-sequential information case. In each case, we introduced a practical assumption of the information's structure that led to meaningful comparisons of the cases. In the uncontrolled case, the lower bound for performance is higher than in the non-sequential case due to lost information energy that is spread across irrelevant edges of the graph. In the controlled case, if the information structure is sufficiently restricted, the lower bound for performance is better than in the non-sequential case.

Chapter 6

The Value of Information in Network Flow Optimization

We now study the impact of partial information in a social welfare context. We consider multiple agents that simultaneously choose decisions from a decision set and where the quality of each decision is dictated by (a) the decision itself, (b) some random perturbation, and (c) the number of agents that choose that same decision.

We motivate this problem with an example. Consider a traffic system where each vehicle corresponds to an agent. The decision set in this case is the set of available paths, and the quality of a decision is the total time of travel. However, the delay of a path is partially dictated by the number of agents who choose to take that path. Therefore, the agents must coordinate themselves in a non-greedy fashion if they wish to maximize their social welfare. We seek to determine how side information can benefit social welfare.

We begin our presentation with a specialization of the general framework in Chapter 2 to social welfare optimization. We then study this framework in the context of a specific problem: network flow optimization. For this application, we provide a specific relaxation that will allow us to express information optimization (on a lower performance bound) as a linear program as well as allow us to derive an analytic performance lower bound.

6.1 Partial Information in Stochastic Social Welfare Optimization

6.1.1 Performance

We first define a framework for studying the value of information in stochastic social welfare optimization.

Let $R \geq 1$ be the number of agents (an integer). Each agent makes a decision $x \in X$. For ease, we assume that X is finite. Let f_x be the fraction of the R agents who choose decision x. Define the vector of decisions as $f = [f_1 \dots f_{|X|}]^T$. Clearly, Rf_x is the number of agents who chose decision x, and Rf is the vector of how many agents chose each decision.

Finally, let $h(x, Rf, W)$ be the quality of decision x. We let the quality be a function of the collective decisions of the agents (the vector Rf) as well as some random perturbation W.

Denote the simplex in $\Re^n$ as $\Delta^n = \{x\Re^n \mid x \geq 0, \|x\|_1 = 1\}$. Clearly, $f \in \Delta^{|X|}$, though by the assumption that R is an integer, f lies in a strict subset of the simplex.

Remark 14. *One can almost immediately see that this framework is actually a specialization of the stochastic optimization in Chapter 2. We have only extended our decision set to the set of all the agents' decisions.*

We assume that every agent has the same information Y about W. For a given Y, the agents optimize the average social performance

$$\min_f \mathrm{E}\left[\sum_{x\in X} f_x h(x, Rf, W) \mid Y\right] = \min_f \sum_{x\in X} f_x \,\mathrm{E}\left[h(x, Rf, W) \mid Y\right].$$

We denote the average social performance under the joint distribution p_{WY} as $J_R(p_{WY})$. This is simply given by

$$J_R(p_{WY}) = \mathrm{E}\left[\min_f \left\{\sum_{x\in X} f_x \,\mathrm{E}\left[h(x, Rf, W)|Y\right]\right\}\right]. \tag{6.1}$$

6.1.2 Quantifying Information and Information Optimization

Finally, given a family of information constraint sets $\{\Gamma(C)\}$, the optimal performance under a capacity C is written as

$$J_R(C) = \min_{p_{WY}\in\Gamma(C)} \{J_R(p_{WY})\}. \tag{6.2}$$

6.2 Network Flow Optimization under Limited Information

In this section, we specialize the general stochastic social welfare framework to network flow optimization.

6.2.1 Specialization to Network Flow Optimization

In network flow optimization, we assume that the agents simultaneously traverse the graph, and that the edge weights of the graph are impacted by the number of agents on each edge as well as some inherent randomness.

The decision set is the set of paths P. Each agent chooses a path $p \in P$ to traverse. f_p is the fraction of agents who take path p, and, as a vector, $f \in \Delta^{|P|}$. We call f the *flow*. The quality $h(p, Rf, W)$ of path p is the total weight of that path, which is sum of

the weights of the edges along that path, and it is defined as

$$h(p, Rf, W) = \sum_{e \in p} \left(\alpha_e R f'_e + W_e\right)$$

where

- $f'_e = \sum_{p \mid e \in p} f_p$ is the fraction of agents that specifically use edge e,
- Rf'_e is the number of agents that use edge e,
- $\alpha_e \geq 0$ is the sensitivity of the edge e's weight to the number of agents using it,
- and W_e is an additive random weight for edge e.

We have some information Y about the edge weights W. The estimate of h given Y is

$$\mathrm{E}\left[h(p, Rf, W)|Y\right] = \sum_{e \in p} \left(\alpha_e R f'_e + \hat{W}_e\right)$$

where $\hat{W} = \mathrm{E}\left[W|Y\right]$.

For a given Y, the agents choose the flow that optimizes the social welfare according to

$$\min_f \left\{ \sum_p f_p \, \mathrm{E}\left[h(p, Rf, W)|Y\right] \right\}.$$

The average performance of the agents under the joint distribution p_{WY} is now just

$$J_R(p_{\hat{W}}) = J_R(p_{WY}) = \mathrm{E}\left[\min_f \left\{ \sum_p f_p \sum_{e \in p} \left[\alpha_e R f'_e + \hat{W}_e\right] \right\} \right]. \tag{6.3}$$

Note that we can denote our distribution p_{WY} as $p_{\hat{W}}$ since $\hat{W}$ is the only random quantity in the expression.

6.2.2 Information Constraints

We define our information constraint sets $\{\Gamma(C)\}$ using the same definition in Chapter 2 for shortest path optimization:

$$\Gamma(C) = \left\{ p_{WY} \mid \mathrm{VAR}\left[\hat{W}\right] \leq C \right\}. \tag{6.4}$$

6.3 Outer Approximating Network Flow Optimization

A key step in generating bounds for J_R is to relax the constraints on the flow f. A similar approach is taken in Chapter 4 for determining the value of information in shortest path

optimization. In Chapter 4, the constraint set for the agent is the set of paths P. By relaxing this constraint set, performance bounds could be derived. A significant difference between network flow optimization and shortest path optimization, however, is that shortest path optimization is a linear optimization while network flow optimization is a quadratic optimization. The methods used in Chapter 4 very much rely on its linear structure, so they are not useful to us here. To this end, we will develop a different approach that leverages the structure of the underlying formulation. Our method requires a minor assumption on the graph's structure.

Definition 3. *An edge cut across a graph is a collection of edges* $\{e_1, e_2, \ldots\}$ *such that each path in the graph passes through exactly one edge in the collection.*

Assumption 3. *There is a subset* $\mathcal{S} = \{S_i\}$ *of all possible sets of edge cuts across* G *such that (a) each path* p *contains an edge from every cut* $S_i \in \mathcal{S}$*, and (b) for two different* $S_i, S_j \in \mathcal{S}$*,* S_i *and* S_j *are disjoint.*

For ease, let $N = |\{S_i\}|$. We expand the set of decisions for each agent using these sets.

Proposition 19. $P \subset \overline{P} = S_1 \times S_2 \times \cdots \times S_N$.

Proof. By Assumption 3 and acyclicity, every path p has exactly one edge in each S_i. □

Remark 15. *In the shortest path problem, the set of decisions for the agent is the set of paths* P *of the graph. A useful outer approximation for* P *is the unit hypercube in* $\Re^{|P|}$*, which is the convex hull of all* 1-0 *combinations of edges. The hypercube approximation effectively expands the set of paths to all possible paths constructed from all the edges (including a path of zero edges that always has zero length). The outer approximation* $\overline{P}$ *that we take for network flow optimization is similar – it is all combinations of edges subject to having only one edge from each cut.*

Proposition 20.

$$J_R(p_{\hat{W}}) \geq \sum_i \mathrm{E}\left[\min_{f \in \Delta_i}\left\{\sum_{e \in S_i} f_e\left[\alpha_e R f_e + \hat{W}_e\right]\right\}\right]$$

where $\Delta_i = \Delta^{|S_i|}$ *and* $f \in \Delta_i$ *is a vector* $f = [f_{e_1}\ f_{e_2}\ \ldots\ f_{e_{|S_i|}}]^T$ *where* $e_k \in S_i$.

Proof. The set of decisions $\overline{P}$ corresponds to a graph, thus we can write the performance of the network flow formulation for this new graph as a lower bound for $J(p_{\hat{W}})$.

$$J_R(p_{\hat{W}}) \geq \mathrm{E}\left[\min_{f}\left\{\sum_{p \in \overline{P}} f_p \sum_{e \in p}\left[\alpha_e R f'_e + \hat{W}_e\right]\right\}\right].$$

Clearly, $f \in \Delta^{|\overline{P}|}$. We can lower bound the optimization further if we allow f to take any value in this simplex (this is akin to making R a real number).

$$\geq \mathrm{E}\left[\min_{f \in \Delta^{|\overline{P}|}} \left\{ \sum_{p \in \overline{P}} f_p \sum_{e \in p} \left[\alpha_e R f'_e + \hat{W}_e \right] \right\}\right].$$

Now we reverse summations.

$$\begin{aligned}
&\geq \mathrm{E}\left[\min_{f \in \Delta^{|\overline{P}|}} \left\{ \sum_{e} \left(\left[\alpha_e R f'_e + \hat{W}_e \right] \sum_{p \mid e \in p} f_P \right) \right\}\right] \\
&= \mathrm{E}\left[\min_{f \in \Delta^{|\overline{P}|}} \left\{ \sum_{e} \left[\alpha_e R f'_e + \hat{W}_e \right] f'_e \right\}\right] \\
&= \mathrm{E}\left[\min_{f \in \Delta^{|\overline{P}|}} \left\{ \sum_{i} \sum_{e \in S_i} \left[\alpha_e R f'_e + \hat{W}_e \right] f'_e \right\}\right]
\end{aligned}$$

We can equivalently optimize over f'_e directly, so we only need to determine the range of its value. Define the vector $f_i \in \Re^{|S_i|}$ as $(f_i)_e = f'_e$ for $e \in S_i$. f_i is simply the vector of flow on edges in cut S_i. Because all flow must pass through edge cut S_i, it is clear that $f_i \in \Delta_i$, yielding

$$\begin{aligned}
&= \mathrm{E}\left[\min_{\{f_i\} \in \Delta_i} \left\{ \sum_{i} \sum_{e \in S_i} \left[\alpha_e R (f_i)_e + \hat{W}_e \right] (f_i)_e \right\}\right] \\
&= \mathrm{E}\left[\sum_{i} \min_{f_i \in \Delta_i} \left\{ \sum_{e \in S_i} \left[\alpha_e R (f_i)_e + \hat{W}_e \right] (f_i)_e \right\}\right] \\
&= \mathrm{E}\left[\sum_{i} \min_{f \in \Delta_i} \left\{ \sum_{e \in S_i} \left[\alpha_e R f_e + \hat{W}_e \right] f_e \right\}\right]
\end{aligned}$$

□

6.4 Information Optimization via Lower Bound Optimization

We now derive computational bounds for performance that are used to provide a rudimentary information optimization algorithm.

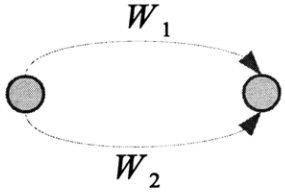


Figure 6-1: Graph for Example 15

6.4.1 The Impact of Information Selection: Examples

Before deriving our first set of bounds, we first provide some examples that highlight how information can impact performance.

Example 15. *Consider the graph in Figure 6-1. Set* $\alpha_1 = \alpha_2 = \alpha$ *and let* $W_i \sim N(0,1)$ *and independent. We consider* $p_{\hat{W}} \subset \Gamma(C)$ *for* $C \leq 2$ *subject to the additional restriction that* $\hat{W}_1$ *and* $\hat{W}_2$ *are independent Gaussians as well. We write* $\hat{W}_i \sim N(0, \lambda_i^2)$.

Evaluating (6.3) *is a matter of solving for the optimal flows* f_1 *and* f_2 *corresponding to each of the two paths. It is straightforward, albeit lengthy, to compute* $J(p_{\hat{W}})$ *for this case, so we only provide the final expression for it:*

$$J_R(p_{\hat{W}}) = J_R(0) + \mathrm{E}\left[\frac{C^2}{\alpha R}[Z]_{\frac{\alpha R}{C}}^{\frac{\alpha R}{C}}\left[[Z]_{\frac{\alpha R}{C}}^{\frac{\alpha R}{C}} - Z\right]\right]$$

where $Z = \frac{\hat{W}_2 - \hat{W}_1}{C}$ [1]. *Unfortunately, even in this simple case, the average performance expression is far too difficult to exactly analyze. With some effort, though, we can lower bound as*

$$J_R(C) \geq J_R(0) + 2C\phi\left(\frac{\alpha R}{C}\right) - \frac{C^2}{\alpha R},$$

where ϕ *is the density of the normal distribution. Though simpler, this expression is not much of an improvement for the purposes of developing a general bound.*

Example 16. *In this example, we compare the performances of two request schemes to motivate the benefit of information optimization. Consider the graph in Figure 6-2 that has* n *disjoint paths with each path having* n *edges. We set* $\alpha_{ij} = 1$ *for all* i, j, *and let* $W_{ij} \sim N(0,1)$ *and independent RVs.*

Define the distribution $p_n^s \in \Gamma(N)$ *that yields* $\hat{W}_{1j} = W_{ij}$ *and* $\hat{W}_{ij} = 0$ *for* $i > 1$ *(serial information). Define the distribution* $p_n^p \in \Gamma(N)$ *that yields* $\hat{W}_{i1} = W_{i1}$ *and* $\hat{W}_{ij} = 0$ *for* $j > 1$ *(parallel information).*

[1] Note that $J(p_{\hat{W}})$ is symmetric with respect to Z and $-Z$, as it should be since Z and $-Z$ have the same distribution.

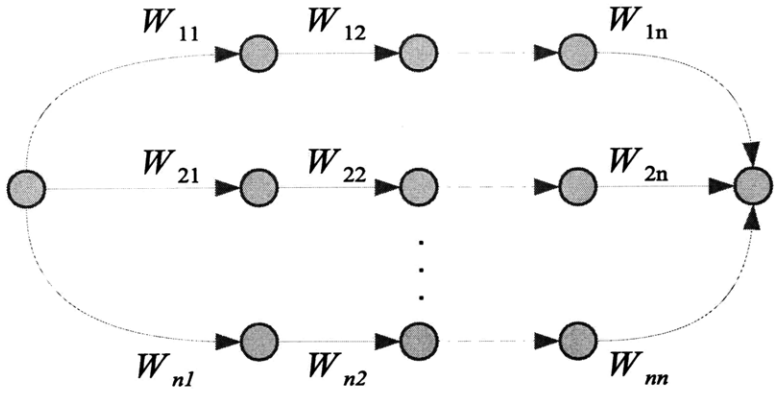


Figure 6-2: Graph for Example 16

The average performances in each case can be written as

$$\begin{aligned} J_R(p_n^s) &= \mathrm{E}\left[\min_f \left\{ \left(Rf_1^2 + f_1 \sum_{j=1}^n \hat{W}_{1j} \right) + \sum_{i=2}^n Rf_i^2 \right\}\right] \\ &= \mathrm{E}\left[\min_f \left\{ f_1\sqrt{n}Z + R\sum_{i=i}^n f_i^2 \right\}\right] \\ J_R(p_n^p) &= \mathrm{E}\left[\min_f \left\{ \sum_{i=1}^n Rf_i^2 + \sum_{i=1}^n f_i\hat{W}_{i1} \right\}\right] \end{aligned}$$

where $Z \sim N(0,1)$. *It is not clear how to exactly evaluate these expressions, so we instead compare bounds for them instead. We hypothesize that the use of* p^s *will (at least asymptotically) outperform* p^p, *so we compare an upper bound for* $J_R(p_n^s)$ *to a lower bound for* $J_R(p_n^p)$.

We derive an upper bound for $J_R(p_n^s)$ *using the selection of a suboptimal flow* f *as well as the inequality* $\|f\|_2^2 \leq 1$. *Since* f *will be a function of* Z, *we choose (the suboptimal)* f *according to* $f_1(Z) = 1$ *if* $Z < 0$ *and* $f_1(Z) = 0$ *if* $Z > 0$. *Therefore,*

$$J_R(p_n^s) \leq \mathrm{E}\left[\min_f \left\{ f_1\sqrt{n}Z + R \right\}\right] = R + \mathrm{E}\left[\min\left\{0, \sqrt{n}Z\right\}\right] = R - \sqrt{n}\frac{1}{\sqrt{2\pi}}.$$

We derive a lower bound $J_R(p_n^p)$ *by simply splitting the* min *operator and using the lower bound* $\|f\| > 0$.

$$\begin{aligned} J_R(p_n^p) &\geq \mathrm{E}\left[\min_f \left\{ \sum_{i=1}^n Rf_i^2 \right\}\right] + \mathrm{E}\left[\min_f \left\{ \sum_{i=1}^n f_i\hat{W}_{i1} \right\}\right] \\ &\geq 0 + \mathrm{E}\left[\min_i \left\{ \hat{W}_{i1} \right\}\right] \end{aligned}$$

$$\geq -\sqrt{2\ln n}$$

where the last inequality is obtained by using Lemma 3 in [8].

Therefore, $J_R(p_n^s)$ (serial information) asymptotically outperforms $J_R(p_n^p)$ (parallel information).

6.4.2 Computational Performance Bounds

We now seek to derive a computational lower bound for performance that is amenable to information optimization. We begin with the follow lemma, which allows us to remove the effect of mean μ from our bounds.

Lemma 5.

$$J_R(p_{\hat{W}}) \geq J(0) - \sum_i \mathrm{E}\left[\min_{f\in\Delta_i}\left\{\sum_{e\in S_i} f_e\left[\alpha_e R f_e + (\hat{W}_e - \mu_e)\right]\right\}\right]$$

where $J(0) = \min_{p\in P}\{p^T\mu\}$ and Δ_i and f are defined as in Proposition 20.

Proof. Adding and subtracting μ and using the relationship

$$\min_x\{a(x)+b(x)\} \geq \min_x\{a(x)\} + \min_x\{b(x)\}$$

yields

$$\begin{aligned} J_R(p_{\hat{W}}) \geq & \mathrm{E}\left[\min_{f\in\Delta}\left\{\sum_p f_p \sum_{e\in p}\hat{W}_e\right\}\right] \\ & + \mathrm{E}\left[\min_{f\in\Delta}\left\{\sum_p f_p \sum_{e\in p}\left[\alpha_e R f'_e + (\hat{W}_e - \mu)\right]\right\}\right] \\ = & \min_{f\in\Delta}\left\{\sum_p f_p \sum_{e\in p}\mu_e\right\} + \mathrm{E}\left[\min_{f\in\Delta}\left\{\sum_p f_p \sum_{e\in p}\left[\alpha_e R f'_e + (\hat{W}_e - \mu)\right]\right\}\right] \end{aligned}$$

The first optimization can be interpreted as a congestionless optimization, the minimizing flow for which is the one that has all agents taking the shortest path in the graph. Therefore

$$\geq J(0) + \mathrm{E}\left[\min_{f}\left\{\sum_p f_p \sum_{e\in p}\left[\alpha_e R f'_e + (\hat{W}_e - \mu)\right]\right\}\right].$$

Bounding the right-hand term using Proposition 20 proves the claim. □

We now present the computational lower bound that is the basis for our information optimization algorithm.

Proposition 21.

$$J_R(p_{\hat{W}}) \geq J(0) - \frac{1}{4R}\sum_e \frac{\mathrm{VAR}\left[\hat{W}_e\right]}{\alpha_e} - \sum_i \frac{1}{\sqrt{\sum_{e\in S_i}\alpha_e^{-1}}}\sqrt{\sum_{e\in S_i}\frac{\mathrm{VAR}\left[\hat{W}_e\right]}{\alpha_e}}.$$

Proof. By Lemma 5, we can compute a lower bound for performance over a single cut and then simply sum over the cuts for the lower bound. Hence, we are interested in computing a lower bound for the expression

$$\min_{f\in\Delta_i}\left\{\sum_e f_e\left[\beta_e f_e + (\hat{W}_e - \mu_e)\right]\right\}$$

where $\beta_e = R\alpha_e$. Assume without loss of generality that $\mu = 0$.

Taking the Lagrangian dual of the optimization, we get

$$= \max_{\gamma,\lambda\geq 0}\left\{\min_f\left\{\frac{f^TQf}{2} + \hat{W}^Tf + \gamma 1^Tf - 1 - \lambda^Tf\right\}\right\},$$

where $Q = \mathrm{diag}\{2\beta_e\}_e$. The optimal f is given by $f = Q^{-1}\left(\lambda - \gamma 1 - \hat{W}\right)$. Substituting this value for f and simplifying, we get

$$= \frac{1}{2}\max_{\gamma,\lambda\geq 0}\left\{-(\lambda - \gamma 1 - \hat{W})^TQ^{-1}(\lambda - \gamma 1 - \hat{W}) - 2\gamma\right\},$$

for which the optimizing γ is $\gamma = \frac{1^TQ^{-1}(\lambda-\hat{W})-1}{1^TQ^{-1}1}$. Substituting this value for γ and simplifying, we get

$$= \frac{1}{2}\max_{\lambda\geq 0}\left\{-(\lambda - \hat{W})^TQ^{-1}(\lambda - \hat{W}) + \frac{1}{1^TQ^{-1}1}\left[(\lambda - \hat{W})^TQ^{-1}11^TQ^{-1}(\lambda - \hat{W}) - 2(\lambda - \hat{W})^TQ^{-1}1 + 1\right]\right\}.$$

We lower bound the expression by removing the positive quadratic term and the positive constant to get

$$= \frac{1}{2}\max_{\lambda\geq 0}\left\{-(\lambda - \hat{W})^TQ^{-1}(\lambda - \hat{W}) - \frac{2}{1^TQ^{-1}1}(\lambda - \hat{W})^TQ^{-1}1\right\}.$$

We achieve a lower bound for the above expression by choosing a suboptimal λ. The unconstrained minimizing value for λ is given by the solution to the equation

$$-2Q^{-1}(\lambda - \hat{W}) - \frac{2}{1^TQ^{-1}1}Q^{-1}1 = 0,$$

for which the solution is $\lambda = -\frac{1}{1^T Q^{-1} 1} + \hat{W}$. However, since λ is constrained to non-negativity, we simply select $\lambda_e = \left[\hat{W}_e\right]^+$. Substituting this value for λ and simplifying, we get

$$\geq -\frac{1}{2}\left[-\left[\hat{W}^T\right]^- Q^{-1}\left[\hat{W}\right]^- + \frac{2}{1^T Q^{-1} 1}\left[\hat{W}^T\right]^- Q^{-1} 1\right].$$

Substituting $Q^{-1} = \frac{1}{2}R^{-1} \cdot \text{diag}\{\alpha_e^{-1}\}_e$, simplifying, and taking expectations, we get

$$= -\frac{1}{4R}\sum_e \frac{\mathrm{E}\left[\left(\left[\hat{W}^T\right]^-\right)^2\right]}{\alpha_e} + \frac{2}{\sum_e \alpha_e^{-1}}\sum_e \mathrm{E}\left[\hat{W}\right]^- \alpha_e^{-1}.$$

Letting $\gamma_e^2 = \text{VAR}\left[\hat{W}_e\right]$, we have the inequalities

$$\mathrm{E}\left[\left(\left[\hat{W}^T\right]^-\right)^2\right] \leq \gamma_e^2$$
$$\mathrm{E}\left[\hat{W}\right]^- \geq -\frac{1}{2}\sqrt{\gamma_e^2}$$

where the second inequality comes from the Cauchy-Schwartz inequality [5]. Thus, we have

$$\geq -\frac{1}{4R}\sum_e \frac{\gamma_e^2}{\alpha_e} - \sum_e \frac{\sqrt{\gamma_e^2}\alpha_e^{-1}}{\sum_e \alpha_e^{-1}}.$$

Applying Jensen's Inequality to the second expression gives

$$\geq -\frac{1}{4R}\sum_e \frac{\gamma_e^2}{\alpha_e} - \frac{1}{\sqrt{\sum_e \alpha_e^{-1}}}\sqrt{\sum_e \frac{\gamma_e^2}{\alpha_e}}.$$

Summing over i (the cuts) yields the bound in the claim. □

6.4.3 Information Optimization

We now present the main result of this section: an information optimization algorithm.

Theorem 11. *A lower bound for $J_R(C)$ is given by the concave maximization*

$$J_R(C) \geq J(0) - \max_{\gamma \geq 0, \|\gamma\|_1 \leq C}\left\{\frac{1}{4R}\sum_e \frac{\gamma_e}{\alpha_e} + \sum_i \frac{1}{\sqrt{\sum_{e\in S_i}\alpha_e^{-1}}}\sqrt{\sum_{e\in S_i}\frac{\gamma_e}{\alpha_e}}\right\}$$

.

Proof. Simply substitute $\gamma_e = \text{VAR}\left[\hat{W}\right]$ in Proposition 21. □

6.5 An Analytic Relationship between Capacity and Performance

Finally, we use Proposition 21 to derive an analytic lower bound for $J_R(C)$ in terms of the capacity C, the total flow R, and certain characteristics of the graph when the α_e's are constant.

Theorem 12. *If $\alpha_e = \alpha$ and $|S_i| \geq M$, then $J_R(C) \geq J(0) - \frac{1}{4\alpha R}C - \sqrt{\frac{N}{M}}\sqrt{C} \sim -O(C)$.*

Proof. Let $\gamma_e^2 = \text{VAR}\left[\hat{W}_e\right]$. If $\alpha_e = \alpha$, then by Proposition 21,

$$\begin{aligned} J_R(p_{\hat{W}}) \geq & J(0) - \frac{1}{4\alpha R}\sum_e \gamma_e^2 - \sum_{i=1}^{N}\sqrt{\frac{\alpha}{|S_i|}}\frac{1}{\sqrt{\alpha}}\sqrt{\sum_{e\in S_i}\gamma_e^2} \\ \geq & J(0) - \frac{1}{4\alpha R}\sum_e \gamma_e^2 - \sqrt{\frac{1}{M}}\sum_{i=1}^{N}\sqrt{\sum_{e\in S_i}\gamma_e^2}. \end{aligned}$$

Using Jensen's Inequality to add the square root terms:

$$\sum_{i=1}^{N}\sqrt{X_i} = N\sum_i \frac{\frac{1}{N}\sqrt{X_i}}{\sum_i \frac{1}{N}} \leq N\sqrt{\sum_i \frac{\frac{1}{N}X_i}{\sum_i \frac{1}{N}}} = \sqrt{N}\sqrt{\sum_i X_i}$$

and using the bound $\sum_e \gamma_e^2 \leq C$ yields the bound in the claim. □

Remark 16. *In shifting the effect of μ into $J(0)$, we get $J_R(0) \geq J(0)$. In general, the two are not equal. Therefore, this bound is necessarily offset from the true performance for any graph where $J_R(0) \neq J(0)$.*

We compare our analytic bound to the actual performance of the optimization for a simple example.

Example 17. *Once again, consider the graph in Figure 6-1. Set $\alpha_1 = \alpha_2 = 1$ and let $W_i \sim N(0, 10)$ and independent. We consider $p_{\hat{W}} \subset \Gamma(C)$ for $C \leq 20$ subject to the additional restriction that $\hat{W}_1$ and $\hat{W}_2$ are independent Gaussians as well. We assume a continuous flow of agents, allowing us to simply normalize $R = 1$. One can interpret the use of continuous flow as a normalized approximation to having many agents.*

The plot in Figure 6-3 shows the simulated performance of the agents as the capacity is varied in the range $0 \leq C \leq 20$. The capacity is equally allocated between the two edge weights so that $\text{VAR}\left[\hat{W}_i\right] = \sqrt{\frac{C}{2}}$. For each value of C, $J_R(C)$ is approximated by averaging $5,000$ simulations.

Interestingly, the performance seems to grow linearly with capacity, as predicted by the bound in Corollary 12.

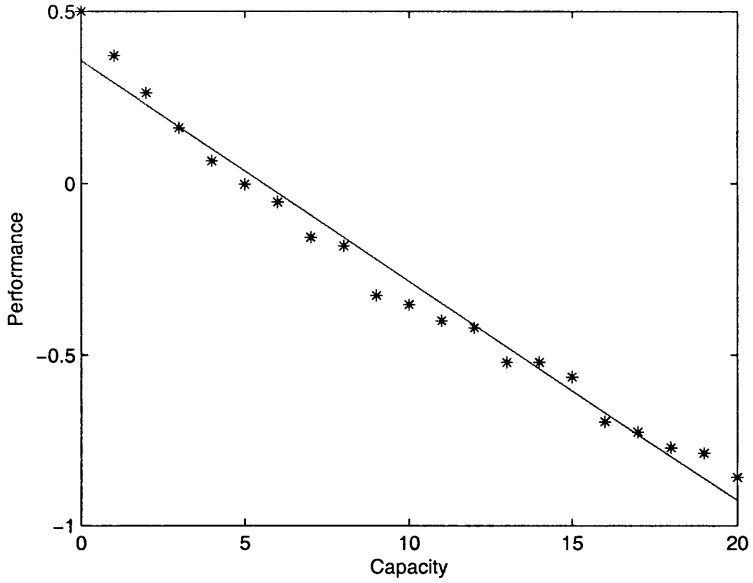


Figure 6-3: Simulated performance of continuous flow with $R = 1$ on the graph in Figure 6-1 as capacity is varied from $0 \leq C \leq 20$. Each point represents a data point from the simulation. The solid line is a linear regression of those points.

6.6 Summary

In this paper, we examined the relationship between information quantity and performance in network flow optimization. It was shown that obtaining analytic bounds for network flow optimization is difficult even in rudimentary cases. Hence, we followed the approach taken Chapter 4 and constructed an outer approximation for the constraint set (based in cuts of the graph) to lower bound the optimization. Both analytic bounds and computational bounds useful for information optimization were developed. The information optimization algorithm in this paper corresponded to a linear optimization that placed more information resources on edges that were more sensitive to congestion. Overall, the fastest rate of improvement with information is linear in the amount of information available to the agents. An example illustrated that linear growth with capacity may be a tight bound on growth.

Chapter 7

Conclusions and Future Work

In this thesis, we examined the value of information in an instance of an uncertain decision framework: shortest path optimization on a graph with random edge weights. In the most basic version of the problem, we considered an agent receiving side information about the edge weights in advance of its travel, using that information to estimate the edge weights, and then traveling along the path with the smallest estimated length. We sought to understand how side information impacted the agent's average performance as well as to understand how the agent should optimize its side information.

We defined a notion of information that was compatible with this formulation: the variance of the edge-weight estimates. By allowing the agent to optimize the information it receives, we needed to develop algorithms for information optimization. A practical set of such algorithms was developed by allowing the agent to optimize bounds for average performance. The bounds were based in a new graph reduction for shortest path optimization that simplified the description of the graph to the minimal-sphere and subspace containing it. In certain cases, the optimization can be practically bounded by a convex optimization.

We also provided a generalization to a sequential-information framework whereby the agent can receive and apply information as it traverses the graph. We considered two cases of sequential information, controlled and uncontrolled. In the controlled case, we allowed the agent to optimize information it would receive at each stage at that stage. In the uncontrolled case, we forced the agent to optimize the information it would receive at each stage before traversing the graph altogether. We defined a new information constraint set for this case (though still based in the variance of the edge-weight estimate) and developed a new abstraction for sequential decision making that greatly simplified analysis. The lower bounds we developed were a generalization of those in the non-sequential case, and the bounds indicated a loss of performance in the uncontrolled-information case while showing a potential gain in performance in the controlled-information case.

Finally, we applied our framework the study the value of information in network flow optimization. It was shown that obtaining analytic performance bounds for network flow optimization is difficult even in simple cases, and, hence, we followed the approach taken in shortest path optimization by constructing an outer approximation for the constraint set

(based in cuts of the graph, not spheres). Both analytic bounds and computational bounds useful for information optimization were developed.

Future work includes further examining the relationship between spheric outer approximation and the edge-cut outer approximations used in Chapters 4 and 6 to actual rates of improvement. The bounds seem to capture the best possible rates of growth, and relating them to optimizations such as that in Corollary 7 may reveal important general structures that can be used to bound more general partial-information optimizations. Future work also includes better understanding the impact of controlled information in the sequential information setting by developing bounds that leverage coarser graph topology information and developing broader conditions under which controlled sequential information can outperform non-sequential information.

Appendix A

Additional Proofs and Results

Chapter 4

Proof of Theorem 2. We prove the result by providing such an algorithm. First, define

$$
\begin{aligned}
f(\{n^1,\ldots,n^i\},j) = \underset{(n,\Delta)}{\operatorname{argmin}} \|n\| \text{ subject to} \\
& n_j = 1,\ \ (n^k)^T n = 0 \text{ for all } k \le i \\
& \overline{J}(v) = \max_{e|\,\mathrm{hd}(e)=v} \left\{\overline{J}(\mathrm{tl}\,(e)) + n_e\right\},\ \ \underline{J}(v) = \min_{e|\,\mathrm{hd}(e)=v} \left\{\underline{J}(\mathrm{tl}\,(e)) + n_e\right\} \\
& \overline{J}(s) = \underline{J}(s),\ \ \overline{J}(t) = \underline{J}(t) = 0
\end{aligned}
$$

with the definition $f(\{n^1,\ldots,n^{i-1}\},j) = (0,\infty)$ if the optimization is infeasible.

The algorithm is as follows:

1: $H_{\mathcal{P}} \leftarrow I$; $i \leftarrow 1$; $j \leftarrow 1$
2: **while** $j \le |E|$ **do**
3: $\quad (n^*,\Delta^*) \leftarrow f(\{n^1,\ldots,n^{i-1}\},j)$
4: $\quad$ **if** $\Delta^* = 0$ **then**
5: $\quad\quad H_{\mathcal{P}} \leftarrow H_{\mathcal{P}} - \frac{n^*(n^*)^T}{\|n^*\|^2}$
6: $\quad\quad n^i \leftarrow n^*$
7: $\quad\quad i \leftarrow i+1$
8: $\quad$ **else**
9: $\quad\quad j \leftarrow j+1$
10: $\quad$ **end if**
11: **end while**

It is clear that the above algorithm terminates in polynomial number of executions, so we only need to prove that it generates a valid projection matrix for $S_{\mathcal{P}}$. For any orthogonal basis $\{n^1,n^2,\ldots,n^k\}$ of $S_{\mathcal{P}}^{\perp}$, a projection matrix for $S_{\mathcal{P}}$ can be written as $I - \sum_{i=1}^{m} \frac{n^i(n^i)^T}{\|n^i\|^2}$. Therefore, we need to show that the set $\{n^i\}$ generated from the algorithm is such a basis.

A vector $n \in \Re^{|E|}$ being normal to $\mathcal{S}_{\mathcal{P}}$ is equivalent to $n^T(\mathcal{P}-\overline{p}) = \{0\} \Leftrightarrow n^T(p-\overline{p}) = 0$

for all $p \in P$. Therefore, a sufficient and necessary condition for n to be normal is that all paths from s to t have the same length when the edge weights are n.

The longest path in the graph when the edge weights are given by n is the unique function $\hat{J} : V \to \Re$ satisfying $\overline{J}(v) = \max_{e|\,\mathrm{hd}(e)=v} \left\{\overline{J}(\mathrm{tl}\,(e)) + n_e\right\}$. The shortest path is the unique function $\underline{J}$ satisfying the same equality with max $\to$ min. An equivalent condition to all paths having the same length is $\overline{J}(s) = \underline{J}(s)$. Therefore, if f is feasible, any optimal solution vector n^* provided by f must be normal to $\mathcal{S}_{\mathcal{P}}$, and, further, it must be non-zero. Therefore, any set of vectors $\{n^1, n^2, \ldots, n^k\}$ must lie in $S_{\mathcal{P}}^{\perp}$.

By the orthogonality constraint in f, any set of vectors $\{n^1, n^2, \ldots, n^k\}$ yielded by the algorithm must be non-zero and orthogonal, hence the set is a subset of an orthogonal basis for $S_{\mathcal{P}}^{\perp}$.

Now, suppose the set $\{n^1, n^2, \ldots, n^k\}$ is a strict subset of a basis. Then there is a non-zero vector n that is orthogonal to each n^j and lies in $S_{\mathcal{P}}^{\perp}$. Let i be the smallest non-zero component of n, and assume without loss of generality that $n_i = 1$ (n can be normalized to produce this). Finally, let $l = \mathrm{argmax}_j\{n^j | n_i^j = 1\}$.

By the definition of l, $f(\{n^1, \ldots, n^l\}, i)$ is infeasible since, otherwise, $n_i^{l+1} = 1$ yielding $l = l + 1$. By the existence of n, though, we know that $f(\{n^1, \ldots, n^l\}, i)$ is feasible since n satisfies all of the constraints of the optimization. The contradiction implies that $\{n^1, n^2, \ldots, n^k\}$ must be an orthogonal basis for $S_{\mathcal{P}}^{\perp}$. Hence, $H_{\mathcal{P}}$ is a projection matrix for $S_{\mathcal{P}}$. □

Proof of Corollary 7. We present a detailed sketch of the proof. First, if we remove the capacity constraint from (4.3) and instead fix the variances $\mathrm{VAR}\left[\hat{W}_e\right] = \lambda_e^2$, get

$$\underline{J}(\{\lambda_e\}) = \min_{p_{\hat{W}}} \left\{ \mathrm{E}\left[\min_p \left\{ p^T \hat{W} \right\} \right] \right\} \text{ subject to}$$
$$\mathrm{E}\left[\hat{W}\right] = \mu,\ \mathrm{VAR}\left[\hat{W}\right] = \lambda_e^2$$

where, for ease, we denote the lower bound for performance in this case as $J(\lambda)$. This optimization is of the form Equation (3.7) in [4]. By Theorem 3.1 in [4], it is equivalent to Equation (3.8) in [4]. Substituting our constraints yields a quadratic objective and second-degree polynomial inequalities, which we can re-express as operations and inequalities on semi-definite matrices:

$$\underline{J}(\{\lambda_e\}) = \max_{\{G_e\}, d \in \Re^{|E|}} \left\{ J(d) + \sum_e G_e \cdot \begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix} \right\} \text{ subject to}$$
$$G_e \leq \left\{ 0, \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -d_e \end{bmatrix} \right\}$$

where by our definition of $J(W)$, $J(d)$ is simply $J(W)$ with $W = d$, a constant.

Let $\Gamma'(C) = \{\lambda \in \Re^{|E|} \mid 0 \leq \lambda_e^2 \leq \sigma_e^2, \|\lambda\|_2^2 \leq C\}$. $\Gamma'(C)$ is clearly a convex set, and any

$p_{\hat{W}} \in \Gamma(C)$ will yield a set of variances VAR $\left[\hat{W}_e\right] = \lambda_e^2$ in $\Gamma'(C)$. We have

$$\underline{J}(C) = \min_{\lambda \in \Gamma'(C)} \underline{J}(\lambda).$$

Taking the dual of the inner optimization for $\underline{J}(\gamma)$ yields a new inner optimization

$$\min_{H_e} \max_{d} \left\{ J(d) - \sum_e H_e \cdot \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -d_e \end{bmatrix} \right\} \text{ subject to}$$
$$H_e \leq 0, \quad H_e \geq - \begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix}.$$

Let $H_e = \begin{bmatrix} a_e & b_e \\ b_e & c_e \end{bmatrix}$. Letting $c = [c_1 \ \ldots \ c_{|E|}]^T$ and $d = [d_1 \ \ldots \ d_{|E|}]^T$, the objective in the minimax is

$$\begin{aligned} &= J(d) + c^T d - \sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e \\ &= \min_{p \in \mathcal{P}} \{p^T d\} + c^T d - \sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e \\ &= \min_{p \in \mathcal{P}} \{(p+c)^T d\} + \sum_e \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \cdot H_e \end{aligned}$$

Of interest to us is the minimaximin expression:

$$\min_c \max_d \min_{p \in \mathcal{P}} \{(p+c)^T d\}$$

If $-c \in \mathcal{P}$, then this expression must always be nonpositive since $0 \in \mathcal{P} - c$. If $-c \notin \mathcal{P}$, then one can show that the expression will always be ∞. Therefore, we require $-c \in \mathcal{P}$. In this case, $d = 0$ is the optimal strategy for d since that maximizes the expression to 0.

The constraint $c \in \mathcal{P}$ is represented by

$$\sum_e \left(\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \cdot H_e \right) v_e \in \mathcal{P}.$$

The remainder of the claim easily follows. □

Chapter 5

Proposition 22.

$$\begin{aligned}\mathrm{E}\left[J_n(Y_n, Y_{\overline{n-1}}, S_{n-1})|Y_{\overline{n-1}}\right] \geq -\sqrt{(r_o^*)^2 - \gamma_{\overline{n-1}}^2}\Big(&\mathrm{E}\left[\hat{W}_n^T H_{n-1}\hat{W}_n|Y_{\overline{n-1}}\right] \\ &+ 2\,\mathrm{E}\left[\hat{W}_n^T H_{n-1}\hat{W}_{\overline{n-1}}|Y_{\overline{n-1}}\right] + \hat{W}_{\overline{n-1}}^T H_{n-1}\hat{W}_{\overline{n-1}}\Big)^{\frac{1}{2}} \\ &+ c_{\overline{n-1}}\hat{W}_{\overline{n-1}}.\end{aligned}$$

Proof. We start with the lower bound from Lemma 4, and we assume without loss of generality that $\mu = 0$ (so that $J(0) = 0$ as well). We first "center" $\overline{\mathcal{P}} \bigcap S$ by removing c_o^*. First, we have

$$\begin{aligned}\underline{J}_1(Y_1) &= J(0) + \min_{S_1 \in \mathcal{S}_1} \left\{\mathrm{E}\left[\underline{J}_2(Y_2, Y_1, S_1)|Y_1\right]\right\} \\ \underline{J}_i(Y_i, Y_{\overline{i-1}}, S_{i-1}) &= \min_{S_i \in \mathcal{S}_i(S_{i-1})} \left\{\mathrm{E}\left[\underline{J}_{i+1}(Y_{i+1}, Y_{\overline{i}}, S_i)|Y_{\overline{i}}\right]\right\} \\ \underline{J}_n(Y_n, Y_{\overline{n-1}}, S_{n-1}) &= \min_{p \in (\overline{\mathcal{P}} - c_o^*) \bigcap S_{n-1}} \left\{p^T \hat{W}_{\overline{n}}\right\} + (c_o^*)^T \hat{W}_{\overline{n}}.\end{aligned}$$

It is straightforward that $\underline{J}_1 = \underline{J}_1'$ where

$$\begin{aligned}\underline{J}_1'(Y_1) &= \min_{S_1 \in \mathcal{S}_1} \left\{\mathrm{E}\left[\underline{J}_2(Y_2, Y_1, S_1)|Y_1\right]\right\} + \mathrm{E}\left[(c_o^*)^T \hat{W}_{\overline{n}}\right] \\ \underline{J}_i'(Y_i, Y_{\overline{i-1}}, S_{i-1}) &= \min_{S_i \in \mathcal{S}_i(S_{i-1})} \left\{\mathrm{E}\left[\underline{J}_{i+1}(Y_{i+1}, Y_{\overline{i}}, S_i)|Y_{\overline{i}}\right]\right\} \\ \underline{J}_n'(Y_n, Y_{\overline{n-1}}, S_{n-1}) &= \min_{p \in (\overline{\mathcal{P}} - c_o^*) \bigcap S_{n-1}} \left\{p^T \hat{W}_{\overline{n}}\right\}.\end{aligned}$$

where $\mathrm{E}\left[(c_o^*)^T \hat{W}_{\overline{n}}\right] = 0$. Therefore, assume without loss of generality that $c_o^* = 0$. Note that this implies that $\overline{\mathcal{P}} = B(0, r_o^*)$.

By definition, $\overline{\mathcal{P}} \bigcap S_{n-1} = B(c_{\overline{n-1}}, r') \bigcap S_{n-1}$ for some $r' \leq r$. For ease, call this intersection B. Since $B \subset S_{n-1}$, $H_{n-1}B = B$. Therefore,

$$\begin{aligned}\underline{J}_n(Y_n, Y_{\overline{n-1}}, S_{n-1}) &= \min_{p \in B} \left\{p^T \hat{W}_{\overline{n}}\right\} \\ &= \min_{p \in (B - c_{\overline{n-1}})} \left\{p^T \hat{W}_{\overline{n}}\right\} + c_{\overline{n-1}}^T \hat{W}_{\overline{n}} \\ &= \min_{p \in (B - c_{\overline{n-1}})} \left\{(H_{n-1}p)^T \hat{W}_{\overline{n}}\right\} + c_{\overline{n-1}}^T \hat{W}_{\overline{n}} \\ &= -r' \|H_{n-1}\hat{W}_{\overline{n}}\| + c_{\overline{n-1}}^T \hat{W}_{\overline{n}}.\end{aligned}$$

By the relationship

$$\mathrm{E}\left[\hat{W}_{\overline{j}}|Y_{\overline{j-1}}\right] = \mathrm{E}\left[\mathrm{E}\left[W|Y_{\overline{j}}\right]|Y_{\overline{j-1}}\right] = \mathrm{E}\left[W|Y_{\overline{j-1}}\right] = \hat{W}_{\overline{j-1}}$$

and Jensen's Inequality, we have

$$\mathrm{E}\left[\underline{J}_n(Y_n, Y_{\overline{n-1}}, S_{n-1})|Y_{\overline{n-1}}\right] = -r'\sqrt{\mathrm{E}\left[\|H_{n-1}\hat{W}_{\overline{n}}\|^2|Y_{\overline{n-1}}\right]} + c_{\overline{n-1}}^T\hat{W}_{\overline{n-1}}.$$

Using the relationship $\hat{W}_{\overline{n}} = \hat{W}_n + \hat{W}_{\overline{n-1}}$ and $H_{n-1}H_{n-1} = H_{n-1}$, we have

$$\|H_{n-1}\hat{W}_{\overline{n}}\|^2 = \hat{W}_n^T H_{n-1}\hat{W}_n + 2\hat{W}_n^T H_{n-1}\hat{W}_{\overline{n-1}} + \hat{W}_{\overline{n-1}}^T H_{n-1}\hat{W}_{\overline{n-1}}.$$

Finally, it is straightforward to show that $(r')^2 = (r_o^*)^2 - \|c_{\overline{n-1}}\|^2 = (r_o^*)^2 - \gamma_{\overline{n-1}}^2$. The bound in the claim follows from the fact that $J_n \geq \underline{J}_n$. □

Appendix B

The Value of Information in Energy Production

We consider an application of the general framework presented in Chapter 2 to energy production whereby the energy supplier seeks to minimize the cost of energy production subject to having some information about (a random) demand. Demand information is important to the supplier in our setting because the supplier is penalized if it undersupplies demand.

Let $D(t)$ be the random demand for power at time t with $\mathrm{E}[D(t)] = \mu(t)$ and $\mathrm{VAR}[D(t)] = \sigma^2(T)$, and let $p(t)$ be the power produced at time t. The supplier's cost $c(p, D)$ of production is a function of both p and D, which, in our simplified setting, is defined as

$$c(p, D) = \sum_t \alpha p(t) + \sum_t \beta [D(t) - p(t)]^+$$

where α is the unit cost of power, and β is the unit cost (in the form of a penalty, such as a fine) of undersupplying demand.

We have some side information Y about D, and we assume that the side information has a certain amount of capacity $C(t)$ at each time t. We define our information set $\Gamma(C)$ similarly as before: $\Gamma(C) = \{p_{DY} \mid \mathrm{VAR}\left[\hat{D}(t)\right] \leq C(t)\}$.

We seek to determine the value of demand information to the supplier. Hence, we seek to bound

$$\begin{aligned} J(C) &= \min_{p_{DY} \in \Gamma(C)} \left\{ \mathrm{E}\left[\min_p \{\mathrm{E}[c(p, D)|Y]\} \right] \right\} \\ &= \min_{p_{DY} \in \Gamma(C)} \left\{ \mathrm{E}\left[\min_p \left\{ \sum_t \alpha p(t) + \sum_t \beta \, \mathrm{E}\left[[D(t) - p(t)]^+ |Y \right] \right\} \right] \right\}. \end{aligned}$$

We upper bound this quantity by choosing a suboptimal value for $p(t)$ and applying

Jensen's Inequality. We first reparameterize the inner optimization using $\hat{p} = p - \hat{D}$.

$$\begin{aligned}
&\mathrm{E}\left[\min_{p}\left\{\sum_{t}\alpha p(t) + \sum_{t}\beta\,\mathrm{E}\left[[D(t) - p(t)]^{+}\,|Y\right]\right\}\right] \\
&= \mathrm{E}\left[\min_{\hat{p}}\left\{\sum_{t}\alpha(\hat{p}(t) + \hat{D}(t)) + \sum_{t}\beta\,\mathrm{E}\left[\left[D(t) - \hat{D}(t) - \hat{p}(t)\right]^{+}|Y\right]\right\}\right].
\end{aligned}$$

Applying Jensen's Inequality yields

$$\leq \min_{\hat{p}}\left\{\mathrm{E}\left[\sum_{t}\alpha(\hat{p}(t) + \hat{D}(t))\right] + \sum_{t}\beta\,\mathrm{E}\left[\left[D(t) - \hat{D}(t) - \hat{p}(t)\right]^{+}\right]\right\}.$$

Finally, choosing a suboptimal solution $\hat{p} = 0$ yields an upper bound strictly in terms of capacity:

$$J(C) \leq \alpha\sum_{t}\mu(t) + \frac{\beta}{2}\sum_{t}\sqrt{\sigma^2(t) - C(t)}$$

where the inequality $\mathrm{E}\left[D(t) - \hat{D}(t)\right]^{+} \leq \frac{1}{2}\sqrt{\mathrm{VAR}\left[D(t) - \hat{D}(t)\right]}$ follows from [4]

Appendix C

Upper Bounds via Splitting and Pruning

C.1 Computing an Upper Bound by Splitting the Graph

We now consider an alternative information optimization that optimizes an upper bound for $J(p_{\hat{W}})$ that better leverages the structure of the graph to reduce complexity. The upper bound is based in Jensen's Inequality, and it applies a dynamic-programming-like approach to computing the average performance backward through the graph, but it requires that the edge weights obey the Gaussian assumptions discussed before.

Theorem 13. *An upper bound $J(C)$ under independent Gaussian edge weights is*

$$J(C) \leq \min_{\{\gamma_e^2\} \in \Gamma(C)} \left\{ \overline{J}(s, \{\gamma_e^2\}) \right\} \tag{C.1}$$

where

$$\overline{J}(v, \{\gamma_e^2\}) = \int j \frac{\partial}{\partial x} \left(\prod_{e \mid \mathrm{hd}(e)=v} \left[1 - \Phi\left(\frac{x - \overline{J}(\mathrm{tl}\,(e)\,, \{\gamma_e^2\}) - \mu_e}{\gamma_e} \right) \right] \right) |_{x=j} \mathrm{dj}$$

and $\overline{J}(t, \{\gamma_e^2\}) = 0$.

Proof. Let $J(v, \hat{W})$ be the length of the shortest path from vertex v to t under edge weights $\hat{W}$:

$$J(v, \hat{W}) = \min_{e \mid \mathrm{hd}(e)=v} \left\{ J(\mathrm{tl}\,(e)\,, \hat{W}) + \hat{W}_e \right\},$$
$$J(t, \hat{W}) = 0.$$

By Jensen's Inequality and acyclicity, we have

$$\mathrm{E}\left[J(v,\hat{W})\right] \leq \mathrm{E}\left[\min_{e|\,\mathrm{hd}(e)=v}\left\{\mathrm{E}\left[J(\mathrm{tl}\,(e)\,,\hat{W})\right]+\hat{W}_e\right\}\right],$$

so that the set of equations

$$\overline{J}(v,\{\gamma_e^2\}) = \mathrm{E}\left[\min_{e|\,\mathrm{hd}(e)=v}\left\{\overline{J}(\mathrm{tl}\,(e)\,,\{\gamma_e^2\})+\hat{W}_e\right\}\right]$$

yields an upper bound for $\mathrm{E}\left[J(v,\hat{W})\right]$.

Now let,

$$\overline{J}(v,\{\gamma_e^2\},\hat{W}) = \min_{e|\,\mathrm{hd}(e)=v}\left\{\overline{J}(\mathrm{tl}\,(e)\,,\{\gamma_e^2\})+\hat{W}_e\right\}$$

so that $\overline{J}(v,\{\gamma_e^2\}) = \mathrm{E}\left[\overline{J}(v,\{\gamma_e^2\},\hat{W})\right]$.

Using

$$P(\hat{W}_e > x) = 1-\Phi\left(\frac{x-\mu_e}{\gamma_e}\right)$$

and $P(\min_i\{X_i\} > x) = \prod_i P(X_i > x)$ for independent RVs $\{X_i\}$ allows us to compute the CDF for the $\overline{J}(v,\{\gamma_e^2\},\hat{W})$. Taking a derivative and integrating yields its expected value:

$$\overline{J}(v,\{\gamma_e^2\}) = \int j\frac{\partial}{\partial x}\left(\prod_{e|\,\mathrm{hd}(e)=v}\left[1-\Phi\left(\frac{x-\overline{J}(\mathrm{tl}\,(e)\,,\{\gamma_e^2\})-\mu_e}{\gamma_e}\right)\right]\right)|_{x=j}\mathrm{dj}$$

□

C.2 Upper Bound on Lost Performance from Pruning the Graph

The information optimizations presented in this thesis are quite manageable, but we can further improve their performances by leveraging a simple fact about real-world graphs: paths of long average length are almost never the shortest path, so information about them can be neglected. We now provide a bound on the performance lost from pruning such paths. In the framework of optimizing over the path polytope, this corresponds to removing certain extreme points from $\mathcal{P}$.

Once again, define the function $\Theta_X(c) = \mathrm{E}\left[\min\{X,c\}\right]$. If $X \sim N(0,1)$, then $\Theta_X(0) = \frac{-1}{\sqrt{2\pi}}$. We use Θ to write a very simple pruning algorithm. Let p^* be the path in the graph with shortest average length. We want to prune the paths p that are often not the shortest path when compared to p^*. Formally, we prune every path p satisfying

$$\mathrm{E}\left[\min\left\{(p^*)^T W, p^T W\right\}\right] - (p^*)^T\mu = \Theta_{(p-p^*)^T W}(0) \approx 0.$$

If $\mathrm{E}\left[p^T W\right]$ is large, we expect that we should be able to prune p, but without additional assumptions on the relationship between the mean $\mathrm{E}\left[p^T W\right]$ and variance $\mathrm{VAR}\left[p^T W\right]$, it may be the case that p is not pruned, even for very large mean.

C.2.1 Efficient Pruning in the Gaussian Case

We seek to establish easily-verifiable conditions under which can efficiently prune paths as well as compute (or estimate) the loss in performance incurred from the pruning. Although the results in this section are specific to the Gaussian case using the information set $\Gamma_G(C)$, they are very easily generalizable to the case of general distributions.

- $\lim_{c\to\infty} \Theta_X(c) = \mathrm{E}[X]$,
- $\lim_{c\to-\infty} \Theta_X(C) = c$,
- and $\frac{\partial \Theta_X}{\partial c}(c) = 1 - F(c)$ with F being the CDF for X.

We can also show that, under certain conditions, as the variance of X increases, the value of $\Theta_X(c)$ decreases.

Proposition 23. *Let $Z \sim N(0,1)$, and let $X_1 \overset{\mathrm{d}}{=} \sigma_1 Z + \mu$ and $X_2 \overset{\mathrm{d}}{=} \sigma_2 Z + \mu$. If $c \leq \mu$ and $\sigma_1 \leq \sigma_2$, then $\Theta_{X_2}(c) \leq \Theta_{X_1}(c)$.*

Proof. First, notice that

$$\Theta_{X_i}(c) = \mu + \sigma_i \Theta\left(\frac{c-\mu}{\sigma_i}\right).$$

Now, taking a derivative with respect to σ, we get

$$\frac{\partial}{\partial\sigma}\left\{\mu + \sigma\Theta\left(\frac{c-\mu}{\sigma}\right)\right\} = \left[\Theta\left(\frac{c-\mu}{\sigma}\right) - \frac{c-\mu}{\sigma}\right] + \left[F\left(\frac{c-\mu}{\sigma}\right) - 1\right] + F\left(\frac{c-\mu}{\sigma}\right)\frac{c-\mu}{\sigma},$$

which is non-positive if $c \leq \mu$. Since

$$\int_{\sigma_1}^{\sigma_2} \frac{\partial}{\partial\sigma}\left(\mu + \sigma\Theta\left(\frac{c-\mu}{\sigma}\right)\right) = \Theta_{X_2}(c) - \Theta_{X_1}(c),$$

the difference is non-positive if the integrand is non-positive. □

We can now we establish a useful relationship between the mean and variance of a path that will guarantee that as the mean of the path increases, the path's performance relative to $\overline{p}$ becomes inconsequential.

Proposition 24. *Let $Z \sim N(0,1)$, and let $\{X_\mu\}_\mu$ be a family of RVs indexed by μ with $X_\mu \overset{\mathrm{d}}{=} \sigma(\mu)Z + \mu$. If $\sigma^2(\mu) \leq \mu$, then*

$$\lim_{\mu\to\infty} \Theta_{X_\mu}(c) = c.$$

Proof. Assume $\mu \geq c$. By the assumption that $\sigma_\mu \leq \sqrt{\mu}$ and Proposition 23, we have

$$c \geq \Theta_{X_\mu}(c) = \Theta_{\sigma(\mu)Z+\mu}(c) \geq \Theta_{\sqrt{\mu}Z+\mu}(c),$$

which implies

$$|c - \Theta_{\sqrt{\mu}Z+\mu}(c)| \geq |c - \Theta_{X_\mu}(c)|.$$

Therefore, we only need to prove convergence of the lefthand term.

Now,

$$\lim_{\mu\to\infty} c - \Theta_{\sqrt{\mu}Z+\mu}(c) = \lim_{\mu\to\infty} \sqrt{\mu}\left(\frac{c-\mu}{\sqrt{\mu}} - \Theta\left(\frac{c-\mu}{\sqrt{\mu}}\right)\right).$$

As $\mu \to \infty$, the inner term goes to 0, and so we apply L'Hopital's rule to get

$$= \lim_{\mu\to\infty} \Phi\left(\frac{c-\mu}{\sqrt{\mu}}\right)(1-2\mu).$$

Applying L'Hopital's rule once again yields

$$= \lim_{\mu\to\infty} \phi\left(\frac{c-\mu}{\sqrt{\mu}}\right)\left[\frac{1}{2}\mu^{\frac{3}{2}} - c\sqrt{\mu}\right].$$

Since $\phi(x)x^{\frac{3}{2}} \to 0$, the limit converges to 0. □

Verifying that $\mathrm{E}\left[p^T W\right]$ and $\mathrm{VAR}\left[p^T W\right]$ satisfy the conditions of Lemma 24 for every path p can be cumbersome in general, but there is an easily-verifiable sufficient condition for it.

Proposition 25. *If each edge weight W_e in G satisfies $\sigma_e \leq \sqrt{\mu_e}$, then for each path P in G, $\sigma_P \leq \sqrt{\mu_P}$.*

Proof. For a path P, we have

$$\sigma_P = \sqrt{\sum_{e\in P} \sigma_e^2} \leq \sqrt{\sum_{e\in P} \mu_e} = \sqrt{\mu_P}.$$

□

Remark 17. *Neither Proposition 23, Lemma 24, nor Proposition 25 really rely on $Z \sim N(0,1)$, but rather that $\mathrm{E}[Z] = 0$ and $\mathrm{VAR}[Z] = 1$. The Gaussian restriction in this section is because are assuming $p_{\hat{W}} \in \Gamma_G(C)$.*

We can use Proposition 25 to derive a simple, efficient pruning algorithm:

1: Choose some threshold length L.
2: For each vertex $v \in V$, compute the average shortest path length $\underline{J}(v)$ from v to t.
3: Let $\tilde{V} = \{v \in V \text{ such that } \underline{J}(v) \leq L\}$.
4: Let $\tilde{E} = \{e \in E \text{ such that } \mathrm{hd}(e), \mathrm{tl}(e) \in \tilde{V}\}$.
5: Construct a new graph $\tilde{G} = (\tilde{V}, \tilde{E})$.

C.2.2 Performance Loss from Path Pruning

The following theorem provides an upper bound for the performance lost from pruning.

Theorem 14. *Let* $\mathcal{P} = \overline{\mathcal{P}} \bigcup \underline{\mathcal{P}}$ *where* $\overline{\mathcal{P}} = \{p | \mathrm{E}\left[p^T W\right] \geq L\}$ *and* $\underline{\mathcal{P}} = \{p | \mathrm{E}\left[p^T W\right] < L\}$. *Then for any* $p_{\hat{W}} \in \Gamma_G(C)$,

$$0 \leq \mathrm{E}\left[\min_{p \in \underline{\mathcal{P}}} \left\{p^T \hat{W}\right\}\right] - J(p_{\hat{W}}) \leq \sum_{k>L} \left[(\#\ \textit{paths of length}\ k)\left((\overline{L} - k) - \sqrt{k}\Theta\left(\frac{\overline{L}-k}{\sqrt{k}}\right)\right)\right],$$

where $\overline{L} = \overline{p}^T \mathrm{E}\left[W\right]$ *and* $\lim_{k\to\infty}\left((\overline{L} - k) - \sqrt{k}\Theta\left(\frac{\overline{L}-k}{\sqrt{k}}\right)\right) = 0$.

Proof. The left inequality is clear, so we proceed to prove the right inequality. First,

$$\min_{p \in \mathcal{P}} \left\{p^T \hat{W}\right\} = \min_{p \in \overline{\mathcal{P}} \bigcup \underline{\mathcal{P}}} \left\{p^T \hat{W}\right\} = \overline{p}^T \hat{W} + \min_{p \in \overline{\mathcal{P}} \bigcup \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}\right\}.$$

Since $0 \in \{(p - \overline{p})^T W\}$ (take $p = \overline{p}$),

$$\begin{aligned}
&= \overline{p}^T \hat{W} + \min\left\{\min_{p \in \overline{\mathcal{P}} \bigcup \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}\right\} \\
&= \overline{p}^T \hat{W} + \min\left\{\min\left\{\min_{p \in \overline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}\right\}, \min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}\right\} \\
&\geq \overline{p}^T \hat{W} + \min\left\{\min_{p \in \overline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}\right\} + \min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\} \\
&= \overline{p}^T \hat{W} + \min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\} + \min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}\right\},
\end{aligned}$$

where the inequality comes from the fact that $\min\{a, b\} \geq a + b$ for $a, b \leq 0$, and the last equality comes from the fact that $\overline{p} \in \underline{\mathcal{P}}$.

Therefore,

$$\begin{aligned}
&\min_{\underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}\right\} - \min_{p \in \mathcal{P}} \left\{p^T \hat{W}\right\} \leq -\overline{p}^T \hat{W} - \min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\} \\
\Rightarrow &\min_{\underline{\mathcal{P}}} \left\{p^T \hat{W}\right\} - \min_{p \in \mathcal{P}} \left\{p^T \hat{W}\right\} \leq -\min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}.
\end{aligned}$$

Take expectations of both sides. We can upper bound the right side as

$$-\mathrm{E}\left[\min_{p \in \underline{\mathcal{P}}} \left\{(p - \overline{p})^T \hat{W}, 0\right\}\right] \leq -\sum_p \mathrm{E}\left[\min\left\{(p - \overline{p})^T \hat{W}\right\}, 0\right] = -\sum_p \Theta_{(p-\overline{p})^T \hat{W}}(0).$$

Let $\sigma_p = \sqrt{\mathrm{VAR}\left[(p - \overline{p})^T \hat{W}\right]}$ and let $\mu_p = (p - \overline{p})^T \mathrm{E}\left[W\right]$. By the assumption that $p_{\hat{W}} \in \Gamma_G(C)$, we have $(p - p)^T \hat{W} \stackrel{\mathrm{d}}{=} \sigma_p Z_p + \mu_p$ for $Z_p \sim N(0, 1)$.

By Proposition 23, we can upper bound $-\Theta_{(p-\overline{p})^T\hat{W}}(0)$ by maximizing the variance of $\sigma_p = \sqrt{\mathrm{E}\left[p^T W\right]}$. Therefore, an upper bound for the right hand side is

$$-\sum_p \left[\mu_p + \sqrt{\mathrm{E}\left[p^T W\right]}\Theta\left(\frac{-\mu_p}{\sqrt{\mathrm{E}\left[p^T W\right]}}\right)\right].$$

We now simply group together the paths by their lengths, yielding

$$\sum_{k>L}\left[(\#\text{ paths of length } k)\left((\overline{L}-k) - \sqrt{k}\Theta\left(\frac{\overline{L}-k}{\sqrt{k}}\right)\right)\right].$$

Finally, the last claim concerning the limit to zero is simply a restatement of Proposition 24 □

Remark 18. *While the bound in Theorem 14 may not be practical for general computation, it may be applicable to classes of graphs that possess certain regular structures. Noteworthy in the bound is that although the number of paths of a particular long length may be large, Proposition 24 tells us that as long as the number of such paths does not increase too quickly, the performance loss from ignoring them is bounded.*

Bibliography

[1] Terje Aven. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of applied probability*, 22(3):723–728, 1985.

[2] M. Avriel and A. C. Williams. The value of information and stochastic programming. *Operations Research*, 18(5):947–954, 1970.

[3] Aharon Ben-Tal and Eithan Hochman. Stochastic programs with incomplete information. *Operations Research*, 24(2):336–347, 1976.

[4] Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds. *SIAM J. on Optimization*, 15(1):185–209, 2005.

[5] Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Tight bounds on expected order statistics. *Probab. Eng. Inf. Sci.*, 20(4):667–686, 2006.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[7] Fan Chung and Linyuan Lu. The diameter of sparse random graphs. *Adv. Appl. Math.*, 26(4):257–279, 2001.

[8] Randy Cogill. Randomized load balancing with non-uniform task lengths. In *Forty-Fifth Annual Allerton Conference*, September 2007.

[9] Luc P. Devroye. Inequalities for the completion times of stochastic pert networks. *Mathematics of Operations Research*, 4(4):441–447, 1979.

[10] V. Firchau. Bounds for the expected value of information. *OR Spectrum*, 2(2):65–73, June 1980.

[11] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming (web page and software). *http://stanford.edu/ boyd/cvx*, June 2009.

[12] J. H. Hagstrom. Computational complexity of pert problems. *Networks*, 18:47–56, 1988.

[13] Refael Hassin and Eitan Zemel. On shortest paths in graphs with random weights. *Mathematics of Operations Research*, 10(4):557–564, 1985.

[14] C. C. Huang, I. Vertinsky, and W. T. Ziemba. Sharp bounds on the value of perfect information. *Operations Research*, 25(1):128–139, 1977.

[15] Keiiti Isii. On sharpness of tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

[16] Isaac Meilijson and Arthur Nadas. Convex majorization with an application to the length of critical paths. *J. Appl. Probab.*, 16:671–677, 1979.

[17] Michael Rinehart and Munther A. Dahleh. Shortest path optimization under limited information. In *Proceedings of the 48th IEEE Conference on Decision and Control*, December 2009.

[18] S.K. Walley, H.H. Tan, and A.M. Viterbi. Shortest path cost distribution in random graphs with positive integer edge costs. In *Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 1023–1032, March 1993.

[19] Gideon Weiss. Stochastic bounds on distributions of optimal value functions with applications to pert, network flows and reliability. *Oper. Res.*, 34(4):595–605, 1986.